



RESEARCH
ARTICLE



OPEN
ACCESS



PEER
REVIEWED

Avatarbots and the crisis of democracy: From bot armies to metaverse swarms

Constantin Vică *University of Bucharest*

Radu Uszkai *University of Bucharest*

Alexandru Dancu

University of Bucharest; National University of Physical Education and Sport, Bucharest

Cristina Voinea *University of Oxford*

DOI: <https://doi.org/10.14763/2026.3.2118>

Published: 24 September 2026

Received: 15 November 2025 **Accepted:** 17 July 2026

Funding: This research was funded by the European Union (ERC, avataResponsibility, 101117761). It also received funding from the project Counterfactual Assessment and Valuation for Awareness Architecture – CAVAA (European Commission, EIC 101071178), as well as from a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI - UEFISCDI, project number PN-IV-P8.1-PRE-HE-ORG-2024-0174, funded through the University of Bucharest under Contract No. 66PHE/2024.

Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Licence: This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 License (Germany) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. <https://creativecommons.org/licenses/by/3.0/de/deed.en>

Copyright remains with the author(s).

Citation: Vică, C., Uszkai, R., Dancu, A., & Voinea, C. (2026). Avatarbots and the crisis of democracy: From bot armies to metaverse swarms. *Internet Policy Review*, 15(3). <https://doi.org/10.14763/2026.3.2118>

Keywords: Avatars, Metaverse, Synthetic majorities, Elections, Platform governance

Abstract: In this paper, we argue that immersive virtual reality environments, which we call the metaverse, can reproduce and intensify the risks social media platforms raise for democracies. At the center of our analysis are avatarbots, digital agents that resemble human beings, usually in their speech and behaviour, but sometimes also in their appearance, while operating partly or entirely without continuous direct human control. We examine how AI-powered avatarbots can fabricate synthetic majorities, simulate political agency, and distort the signals on which democratic legitimacy depends. To evaluate these risks, we develop the Principle of Accountable Free Expression (PAFE), structured around three criteria: authenticity, non-domination, and verifiability. These correspond to responsibility for speech, insulation from economic domination, and the verifiable origin of political communication. We use PAFE to analyse the democratic risks raised by avatarbots and to derive design and governance guidelines for immersive environments that aim to function as democratic public spheres.

This paper is part of **Governing the metaverse through the lens of the public sphere**, a special issue of *Internet Policy Review* guest-edited by Annelieke A.M. Mooij and Sunimal Mendis.

Introduction

In November 2024, the Romanian Constitutional Court annulled the presidential elections after intelligence services found evidence of foreign interference in the campaign through social media platforms (Romanian Presidential Administration, 2025). What looked, initially, like a surprising grassroots mobilisation campaign for a fringe candidate almost no one had heard about six months before the elections, turned out to be a coordinated network of bots simulating citizen support.

The Romanian case illustrated once again what extensive research had already documented: the infrastructure of social media platforms can be gamed and users' cognitive biases exploited to push various (political) agendas (Wooley & Howard, 2018; Borgesius et al., 2018; Deibert, 2019; Bradshaw & Howard, 2018; Carpenter et al., 2020; Nguyen, 2020; Mihailov, Voinea, & Vică, 2023; Marin & Vica, 2024; Mooij, 2025). But, as people move to more immersive platforms, techniques for influencing and manipulating people might adapt to these new environments.

In this paper, we argue that immersive virtual reality environments, which we call the metaverse, can reproduce and intensify the risks social media platforms raise for democracies. We use the term 'metaverse' to refer to immersive and interactive computer-generated spaces in which users experience themselves and others as inhabiting a shared virtual world. In Chalmer's words, the metaverse is "a three-dimensional world [where] you can see and hear as if you existed within it" (2022, p. xii). We use the term broadly to include both centrally governed corporate platforms/Web2 platforms, such as Meta Horizon Worlds, as well as more decentralised, blockchain-based, Web3 based environments, like Decentraland.

At the centre of our analysis are avatarbots, digital agents that resemble human beings, usually in their speech and behaviour, but sometimes also in their appearance, while operating partly or entirely without continuous direct human control (Constantinescu, 2026, p. 2059). Even more than social media bots, avatarbots can obscure who is actually speaking and for what reasons, while at the same time being able to pass as independent users. When deployed at scale, they could be used to produce a synthetic majority: the artificial appearance that a broad and independent public has already 'spoken' or mobilised around a candidate or position.

To evaluate the legitimacy of political speech¹ generated by avatarbots in the metaverse, we advance the principle of accountable free expression (PAFE). PAFE is structured around three criteria: authenticity, non-domination, and verifiability. These correspond to responsibility for speech, insulation from economic domination, and the verifiable origin of political communication. We use PAFE to analyse the democratic risks raised by avatarbots and to derive governance principles and recommendations for immersive environments that foster democratic participation and deliberation.

Our argument is prospective rather than descriptive, although avatarbots are already here. For example, political campaigns have already begun experimenting with AI-generated candidate avatars (see AI-Yoon in South Korea), conversational campaign agents (like Ashley in the United States and AI Steve in the United Kingdom), and avatar-based campaigning in virtual environments (see Andrew Yang's campaign event on Zepeto). To our knowledge, there are no documented cases of the use of avatarbots in the metaverse for nefarious political purposes. But it is not unreasonable to assume that these technologies too will be used (like all digital technologies before) as vehicles for influencing people and imposing various political agendas.

The argument unfolds as follows. Section 1 explains how the democratic pathologies associated with social media platforms could be reproduced and intensified in immersive environments. Section 2 develops the Principle of Accountable Free Expression as a normative framework for evaluating synthetic political communication. Section 3 examines the extent to which existing EU legislation addresses concerns related to PAFE and identifies further considerations for the design and governance of immersive political environments. The last section concludes.

Section 1. The metaverse: the next frontier

Digital technologies, whether immersive or not, are increasingly being used to lead covert and coordinated digital campaigns to distort public opinion and to create the impression of apparently spontaneous and widely shared support for various products, positions, or candidates. We take the Romanian example as a case study to unpack how synthetic majorities are created in the first place on social

1. By political speech and communication, more broadly, we understand, following McNair (2018), purposeful communication about politics and it includes: "1. All forms of communication undertaken by politicians and other political actors for the purpose of achieving specific objectives; 2. Communication addressed to these actors by non-politicians such as voters and activists; 3. Communication about these actors and their activities, as contained in news reports, editorials and other forms of media discussion of politics, such as blogs and social media posts" (2018, 4).

media platforms. We then show how the metaverse could extend and amplify this phenomenon and explain how avatarbots could contribute to it.

1.1. From bot armies to synthetic majorities: the Romanian case

In the first round of the November 2024 Romanian presidential race, nationalist outsider Călin Georgescu jumped from about 1 percent in polls to 23 percent of the vote after a covert €1 million influence operation (Funky Citizens 2025, 4). Co-ordinated campaigns used TikTok's recommendation systems to amplify fringe nationalist messages, while Telegram groups distributed synchronised content featuring political impersonators and auto-generated debates. Influencers operated in tandem with bot armies to boost Georgescu's visibility across platforms. Italian and French intelligence services later showed that the interference relied on orchestrated campaigns using algorithmically amplified content, AI-generated commenters, and deepfake political impersonations (AISE, 2025; VIGINUM, 2025).

Because of this, Romania's Constitutional Court (RCC) annulled the election, concluding that the result no longer reflected the popular will because of the strong interference in the electoral campaign. In its decision, RCC highlighted how the asymmetrical exposure that Călin Georgescu benefited from via bot armies made the electoral competition unfair by distorting public deliberation (CCR, 2024). The court's decision hinged on a new and dangerous phenomenon: what we here call the creation of a synthetic majority.

A synthetic majority is a specific outcome of manufactured consensus (see Woolley & Guilbeault, 2017; Woolley & Howard, 2018; Woolley, 2023). Woolley (2023, p. 8) defines 'manufactured consensus' as the strategic creation of the "illusion of popularity" for a message or candidate by exploiting social media platforms' infrastructures. It usually stems from three interrelated practices (2023, p. 55): (1) the use of political bots, sock-puppet accounts (false online identities used for deceptive purposes), and partisan influencers to simulate popularity or opposition; (2) the gaming of recommendation and trend systems to present targeted content as reflecting public opinion; and (3) the recycling of these signals by news media, which further legitimate them as evidence of popular sentiment. It's important to note that not every instance of manufactured consensus produces a synthetic majority. For example, inflated likes or other metrics can create the impression that a person, product, service, or position enjoys support, without necessarily making it appear politically dominant. One could, for example, game social media algorithms to manufacture consensus around a certain weight loss product, positioning it as

highly popular and very effective. But a synthetic majority is more than that, because it simulates public will. Unlike shared recognition or public opinion (which manufactured consensus mimics), public will is “a social system’s shared recognition of a particular problem and resolve to address the situation in a particular way through sustained collective action.” (Raile et al., 2014, p. 111). More precisely, public will involves both people’s belief in the fact that they can do something to make a change, as well as their willingness to commit resources to that cause (Raile et al., 2014, p. 115). So, unlike manufactured consensus, synthetic majorities create the appearance of a public ready to act.

Schroeder et al. (2026) show how malicious AI swarms can intensify precisely this process by deploying large numbers of coordinated but apparently independent AI personas that seed compatible narratives across communities, generate artificial social proof, and thereby manufacture the appearance of broad grassroots agreement. More importantly, such coordination can simulate not merely popularity but an apparently autonomous public will: once the independence of individual voices is covertly replaced by orchestrated synthetic agents, citizens may mistake engineered agreement for a genuine majority prepared to support and act on a political position.

The Romanian presidential campaign perfectly illustrates the emergence of a synthetic majority. First, bot networks and paid influencers filled TikTok and Telegram with pro-Georgescu content, simulating a surge of support that did not exist in reality yet. More and more people jumped on the bandwagon, as money was being poured into inflating social media metrics. Second, the gaming of TikTok’s and other social media platform’s recommendation systems then made this content trend, as if it reflected a spontaneous shift in public will; people had the impression that a lot of their fellow citizens already decided to support and vote for Georgescu (Funky Citizens, 2025; EDMO, 2025). Taken together, these dynamics created what we describe as a synthetic majority: the appearance that the public had already rallied behind Georgescu. Whether and to what extent this perception causally affected voting behaviour is a separate empirical question that we do not attempt to settle here.

Finally, the Romanian case underscores the geopolitical stakes of such manipulation. As the Romanian Presidential Administration (2025) reports, cyber-attacks, political subversion, and disinformation have become central to the Russian Federation’s hybrid warfare strategy in Central and Eastern Europe. Comparable covert operations surfaced in other countries, too: Hungary’s 2022 election saw government-aligned chatbots and micro-targeted ads tilting the field (Human Rights

Watch, 2022, p. 15). The same happened in Germany's 2024 European elections, with AfD deliberately targeting young TikTok users (Gatterman & Tulin, 2025). The Republic of Moldova also managed to investigate and contain networks of Russian bots on Facebook and TikTok which targeted the incumbent president and the pro-EU agenda (Marocico, Mirodan, & Ings, 2025).

In this context, synthetic majorities are not merely passing social media trends, but instruments of geopolitical influence. More precisely, the objective of such synthetic majorities is not simply to persuade citizens of a particular claim or to spread misinformation, but to alter people's perception of what their fellow citizens believe and which political forces are dominant. In this way, synthetic majorities are a way of nudging people to act in ways that serve the interests of various parties.

But while on social media, synthetic majorities are constructed through manipulated metrics of popularity, such as likes, shares, trends, and comments, immersive environments could give this strategy a different form. Political actors could deploy avatarbots that appear to be citizens providing support. The manipulation would therefore work by simulating the apparent size, composition, behaviour, and, more importantly, 'feel' of the public itself.

1.2. The metaverse as an intensification of social media pathologies

Much like Facebook and Twitter were initially framed as horizontal, decentralised tools for free expression (Voinea et al., 2020), the metaverse is marketed as an arena for rich, personal connection (Cheng et. al., 2022). As Bolter and Grusin (2020) interestingly argue, new media repackage the forms and logics of older media under the guise of novelty. The metaverse 'remediates' social media by enabling real-time, embodied presence instead of flat timelines. But, it also reproduces the same extractive infrastructure: headsets and sensors track users' gaze, facial micro-expressions, tone, poise, and even heart rate or non-conscious movements such as pupil dilation (Rosenberg, 2023); all of this data is then commodified, as platform governance is dominated by unaccountable corporate interests (Egliston & Carter, 2021). This vast trove of personal, granular data enables what Rosenberg (2023, 124) calls 'interactive influence' and Floridi (2024, p. 64) 'hyper-persuasion' (or 'hypersuasion'): the adjustment and personalisation of persuasive tactics with the help of AI in real-time based on users' biometric data. In immersive environments influence can be exerted by dynamically manipulating not only the informational content the user sees (like on social media platforms), but also features of the en-

vironment, from the gestures, poise, and tone of the avatar one is interacting with, to crowd density, colors, and looks of the world etc. So, social media platforms determine which information a user encountered, but immersive systems determine the experience within which anything is encountered at all. This means that synthetic agents can calibrate how they address each citizen who encounters them, which, in its turn, can make them and the synthetic majorities they can create even more persuasive than on social media platforms.

This is because on social media users normally encounter others through content; that is, they passively read posts or watch live streams through a two-dimensional interface. In immersive environments, by contrast, users can occupy the same virtual space, talk face to face through avatars, and respond to one another through gaze, gesture, posture, and movement. Haptic technologies can add a further sensory dimension by simulating certain forms of touch and physical contact. As a result, other participants are experienced not simply as profiles producing content, but as human beings in face-to-face interactions.

This experiential dimension matters because people do not form political judgments in isolation from their social context and peers. Research on conformity and bandwagon effects show that perceptions of majority opinion can influence both political attitudes and willingness to express dissent (Nadeau, Cloutier, & Guay, 1993; Schmitt-Beck, 2008; Matthes, Knoll, & von Sikorki, 2017; Ross et al., 2019;). For example, Farjam (2020) found that showing participants polls of other people's choices raised support for the most popular option by an average of 7%, while at the same time drawing support away from both minority and intermediate options. Also, Farjam and Loxbo (2024) found that exposure to polls on contentious policy issues undermined minority positions by converting some opinions and demobilising others, with only a small minority of committed partisans resisting both.

One might be tempted to treat this inclination to give in to what the majority believes as a lapse of rationality. But, as Levy (2022) argues, rational deliberation doesn't take place in isolation of others. We often use higher-order evidence (which is evidence about evidence), to decide whom to trust and when deference is warranted. In simpler terms, because we cannot independently evaluate every claim we encounter, we routinely rely on social cues (such as expertise or credibility) to decide what and whom to believe. Deference is therefore, under the right conditions, part of rational reasoning (Levy, 2022, ch. 3).

But problems arise when the epistemic environment in which this deference takes place is polluted. A polluted epistemic environment, on Levy's account, is one in

which the markers of credibility or expertise are mimicked by actors who want to mislead (Levy, 2022, ch. 5). Synthetic majorities create precisely this kind of problem. They mimic a form of higher-order evidence that people normally use to form political judgments: the convergence on a position or candidate of many independent citizens. What makes such convergence epistemically informative is, at least in part, that the judgments originate independently. A synthetic majority removes this independence, while preserving its appearance.

The resulting harm is more subtle than simply exposing citizens to false claims, like on social media platforms. Synthetic majorities distort the informational environment in which political judgment takes place by making coordinated political speech appear to be independent agreement. So, citizens lose access to reliable higher order evidence that they used in political judgments.

What is more, there is some evidence that for social influence to be effective, we don't even need other human beings. In replications of classic conformity experiments in VR, participants have conformed to majorities composed entirely of virtual agents (Kyriltsias & Michael-Grigoriou, 2018; Kyriltsias et al., 2020). Whether participants believed those agents to be human-controlled or computer-controlled made no significant difference to how much they conformed (Kyriltsias & Michael-Grigoriou, 2018; Kyriltsias et al., 2020).

So, the metaverse intensifies the problems already familiar from social media, where the fabricated support of synthetic majorities is typically encountered as information *about* other people, under the form of various metrics. In immersive environments, synthetic majorities could instead be encountered through the apparent presence and behaviour of other people, making their possible impact even more likely. Two further aspects compound this intensification. First, some of the cues that currently help identify coordinated campaigns on social media platforms are not yet available on most immersive platforms. On social-media, we can look at account histories, funding, follower networks, and other traces of coordination to gain information about who is behind an account or a coordinated campaign. Avatarbots, by contrast, can generate spontaneous speech and behaviour in real time, making it harder to identify who is behind it. Second, whereas conventional bot networks often spread content to large numbers of users, avatarbots could engage individual users directly and adapt their speech and non-verbal behaviour to users' responses in real time. Our claim is therefore not that immersive environments exert stronger political effects than social media, but that their affordances create additional mechanisms through which synthetic social influence could operate.

1.3. Avatarbots

But all the risks mentioned above depend on what kind of avatarbot is ‘doing the speaking’, as they are not a homogenous category. Avatarbots differ in what concerns their autonomy and authorisation (Constantinescu, 2026; Hurshman et. al., 2026). By autonomy we mean the extent to which their behaviour is generated by AI rather than directly controlled by a human. Authorisation refers to whether avatarbots act on behalf of an identifiable principal who takes responsibility for their speech.

We distinguish four main types of avatarbots: digital duplicates, digital doppelgängers, autonomous avatar agents, and coordinated avatar swarms.

TABLE
1:
Taxonomy
of
political
avatarbots
by
autonomy
and
authorisation.

TYPE	AUTHORISATION	AUTONOMY
Digital duplicates	Authorised	Low to medium
Digital doppelgängers	Unauthorised	Low to high
Autonomous avatar agents	Authorised/Unauthorised	High
Coordinated avatar swarms	Authorised/Unauthorised	High (collective)

Digital duplicates are representations of real people created with their participation or permission. A realistic 3D duplicate can be created from photos, videos, or scans of the person, which are used to build and animate a digital model. Voice recordings can also be used to reproduce that person’s voice. Once created, the duplicate may be controlled by the original person or another operator, or it may deliver statements prepared in advance (Whittaker, Weismueller, & Pallant, 2026). Because its speech and behaviour is typically decided and controlled by a human being, this type of avatar generally has low or medium autonomy.

Digital doppelgängers are, essentially, just like digital duplicates, with the only dif-

ference that they are unauthorised, as they reproduce a person's likeness without their consent. This can be done by using video, images, or audio of the target to train an avatarobot that can look and sound like the original person (Hurshman et al., 2026). Their autonomy can range from low, when a human operator directly controls what the avatar says and does, to high, when an AI system generates unscripted speech and behaviour based on data associated with the target.

Autonomous avatar agents are not modeled on any existing person, so they are fully generative entities, operated by AI systems rather than human beings. A language model generates their responses and selects appropriate actions. Then, text-to-speech produces the voice, while lip-synchronisation and animation systems create the corresponding facial and bodily performance. They can generate each statement or movement in real time, although their roles and permitted behaviours are configured by their designers (Li et al., 2026).

Coordinated avatar swarms are large-scale, centrally managed clusters of semi-automated avatars (Schroeder et al., 2025). Swarms can be built from doppelgängers or autonomous agents. A single operator can instantiate hundreds of instances of avatarobots and give each one a slightly different name, face, and voice, as well as feed them all the same talking points, so what shows up in immersive environments could look like independent people arriving at the same conclusion.

The democratic problem posed by avatarobots and the synthetic majorities they can be used to create cannot be reduced to misinformation. Even if every claim expressed by an avatarobot were factually accurate, it would still be problematic if coordinated artificial agents could present themselves as independent citizens and make the speech of a few actors appear to express the will of a broader public. So, what avatarobots and synthetic majorities might break even more than social media platforms already did, is the link between political speech and the agent responsible for it.

Section 2. A normative framework

So how should the political speech produced through avatarobots be evaluated? In this section, we propose the Principle of Accountable Free Expression (PAFE). PAFE requires that speech can be attributable to responsible human or institutional actors, protected from domination through concealed economic or technical power, and traceable to a verifiable origin. We formulate these requirements as three criteria: authenticity, non-domination, and verifiability.

But before we proceed, a few clarifications are needed. In this paper, we are working with the assumption that AI systems do not (yet) possess political agency in their own right. They do not have (welfare) interests, or bear responsibility for the consequences of their speech. Although an avatarobot may generate language and behaviour without continuous human control, this functional autonomy does not amount to moral or political agency.

PAFE therefore applies to political speech produced by or through avatarobots, but it does not treat the avatarobot itself as the ultimately accountable speaker. Responsibility must instead be traced to the human or institutional actors who authorise, design, and deploy, as well as benefit from these technologies. So, the relevant question PAFE addresses is whether the political speech these technologies are used to produce can be attributed to identifiable actors who can be held answerable for it.

Also, PAFE is not limited in principle to immersive environments. The same normative concerns can arise wherever synthetic systems mediate political communication. We focus on the metaverse because embodied synthetic agents make especially visible a broader problem, the possibility of technologically simulating independent political agency itself.

2.1. The principle of accountable free expression

To develop the Principle of Accountable Free Expression, we draw on three political philosophers whose work illuminates different conditions of democratic communication. We begin with Mill because his defence of free expression identifies a basic condition of democratic deliberation: the value of public disagreement depends not merely on the circulation of statements, but on the presence of viewpoints that can be articulated by agents who are answerable for them. We then take inspiration from Walzer to clarify why inequalities of wealth and technical capacity should not determine the distribution of political support. Last but not least, we use O'Neill's work to show why communication must remain attributable if it is to sustain public trust. Taken together, these perspectives provide the basis for PAFE's three criteria: authenticity, non-domination, and verifiability.

Mill anchors free speech in an epistemic wager: only a public contest of *real* opinions allows societies to approximate truth and avoid error. Silencing even a single voice, he warns, denies humanity either the chance to exchange falsehood for truth or the opportunity to grasp truth more vividly by grappling with its negation (Mill, 1859/1989, ch. II). Each speaker is presumed to be an authentic human mind

advancing reasons that others can test, which makes the “collision of adverse opinions” so potent. Where claims are endlessly replicated by code rather than defended by persons, the evidential field tilts: repetition masquerades as justification and the search for truth is overshadowed by volume.

Mill defends free speech on formative grounds too. Liberty of expression is the school of character: citizens cultivate judgement, courage, and individuality only by confronting live dissent, by, as he puts it, “hearing what can be said against them”. But that formative value presupposes interlocutors who *can* be heard and *held accountable*. When botnets or avatarobots flood an immersive space with coordinated messages, they can create what Mill called “the tyranny of public opinion”. Rather than encountering a mix of contesting views, users may be exposed to what looks like unanimous support for a particular candidate or cause. A voter who sees most virtual neighbours cheering Candidate X may feel pressure to conform, without realising that the crowd is largely computational.

Where Mill scrutinises *how* voices originate, Walzer scrutinises *which resources* may enter the political arena. In *Spheres of Justice* he argued that a pluralist society sustains complex equality only when goods from one sphere, e.g., wealth, celebrity, technical prowess, cannot buy dominance in another, especially the political (Walzer, 1983, ch. 1). To prevent domination Walzer calls for blocked exchanges: “there must be things that money cannot buy: votes, offices, jury decisions...” (Waltzer, 1983, p. 100). Bots and avatarobots put pressure on this principle by enabling those with economic or computational capital to generate the appearance of civic participation. A candidate or interest group can, in principle, rent server capacity to deploy large numbers of synthetic agents, making support appear broader or more intense than it is. This risks weakening the insulation between economic and political spheres by turning computational resources into a means of amplifying voice. In such a configuration, what looks like popular backing may increasingly reflect the distribution of money and technical capacity rather than the distribution of citizens’ views.

Finally, Onora O’Neill focuses on what both Mill and Walzer implicitly presuppose: that democratic communication requires speech that is attributable and assessable, not just abundant. “Digital technologies are excellent for *expressing* and *disseminating* content, but not for *receiving* and *judging* content,” she writes, “hence not for communication, or for meeting the ethical and epistemic standards that matter [...] for democracy” (O’Neill, 2020, p. 15). The problem is not simply that online actors may lie, but that digital infrastructures make it hard for people to see who is speaking, with what aims, and under whose control. When the link between

speaker and speech act is opaque, expression ceases to play its part in public reason and becomes a vehicle for covert influence. This opacity is incentivised. As O'Neill shows, digital platforms allow for the intervention in public discourse with minimal accountability (O'Neill, 2020, p. 14). We have moved from a public sphere structured by institutional intermediaries (editors, journalists, broadcasters) to one dominated by algorithmic intermediaries whose loyalties lie with engagement metrics, not civic norms (Lazar, 2024).

O'Neill (2020) also emphasises that speech acts require identifiable, responsible agents if they are to be trustworthy. Anonymity (or pseudonymity) and the complexity of algorithmic intermediaries mean that audiences frequently cannot tell who is speaking or why they are amplifying certain content. O'Neill warns that public trust is eroded when communication becomes untraceable: "Citizens in democracies need to share a common space in which debate and communication takes place, and this is likely to be damaged if anonymous interventions in democratic politics can be undertaken by wealthy individuals (who need not be citizens), or by anonymous corporate and state actors, including hostile state actors." (O'Neill, 2020, p. 17). Digital platforms permit exactly such covert interventions, allowing hidden actors to wield influence without entering ordinary structures of public accountability, such as being answerable to voters or regulators.

Our principle of accountable free expression (PAFE) synthesises these three strands: from Mill, we take the idea that the value of free expression depends on agents expressing views they hold and can, in principle, defend; from Walzer, we take the importance of insulating political agency from domination by other spheres; from O'Neill, we recognize that without traceability, attribution, and communicative trust, speech becomes a form of covert manipulation, not public reason.

PAFE is structured around three main criteria: authenticity, non-domination, and verifiability. Authenticity has in view the relation between speech and speaker (or principal). It doesn't require that every utterance be personally composed or directly controlled by that principal; spokespersons, authorised representatives, and automated systems can communicate authentically when the relationship of representation is appropriately disclosed. Non-domination requires that economic or computational resources should not be covertly converted into disproportionate political influence or into the false appearance of independent civic support. Verifiability concerns the relation between speech and its audience: citizens should be able, in principle, to trace a contribution to its source and recognise it as that source's own. These three criteria work together to provide a normative framework

which we can use to assess the legitimacy of political communication.

2.2. Implications for avatarbots

Bots, avatarbots, and other synthetic agents pose fundamental challenges to democratic discourse not because they can spread misinformation (which is something humans also do) but because they undermine the conditions that make democratic communication possible in the first place.

First, authenticity depends on there being an author, or a principal whose views, interests, and reasons the speech expresses. Avatarbots have no interests of their own, they don't hold the views they express in the sense of being committed to defending them, so they can't be held accountable for having expressed them. What follows is that the avatarbot cannot accomplish the authenticity condition on its own, so the condition must be met further up the chain, by whoever authorises and deploys it. Authenticity therefore fails when there is no principal behind the avatarbot at all, and when a real principal is concealed behind the appearance of (many) independent speakers. Also, avatarbots are problematic because they appear as embodied conversational partners, so they could be more easily mistaken for human beings with legitimate standpoints. An automated post can at least be recognised as "content"; an avatarbot designed to resemble an ordinary citizen could more readily be taken as an equal. Second, non-domination is threatened when computational resources can be converted into the appearance of civic support. Text-based bots already allowed those with money and technical capacity to scale up their presence, but avatarbots extend this logic. In their case, what can be purchased is no longer amplification of a message but the apparent existence of the people delivering it. What appears as a broad base of civic engagement can therefore track the distribution of financial resources, server capacity and engineering effort, rather than the distribution of citizens' views.

Third, verifiability becomes harder to secure because immersive environments offer fewer of the traces on which attribution depends. In text-based environments, there are often at least minimal cues, such as usernames, profiles, posting histories, on the basis of which users can form rough judgements about credibility and intent, even if those cues are imperfect. By contrast, avatarbots can convincingly simulate emotions through voice, gesture, and coordinated behaviour, so users are less likely to perceive them as designed instruments of persuasion. Verifiability requires not only disclosure that an avatar is synthetic, but also information about who authorised, financed, controls, or benefits from it.

One of PAFE's advantages is that it identifies which uses of avatarbots for political communication are objectionable and why. For example, digital political duplicates satisfy PAFE's criteria when they operate as authorised and transparent extensions of identifiable people or institutions. Digital doppelgängers undermine authenticity and verifiability by reproducing a person's likeness without consent and obscuring the actors responsible for the impersonation. Autonomous avatar agents fail all criteria when they are presented as independent political participants rather than as synthetic systems attributable to identifiable deployers. Similarly, when the origins and control over coordinated avatar swarms is concealed and centrally organised communication is made to appear as the independent expression of numerous citizens, PAFE's criteria can't be checked.

The harm that the political speech generated by avatarbots can create therefore lies not merely in the possibility of false statements, but in the simulation of democratic agency and public support. The next section considers how existing regulation and the design of immersive platforms might protect accountable free expression against these forms of synthetic political presence.

Section 3. Democratic design in the age of synthetic agency

Before applying PAFE to the design and governance of metaverse platforms, we need to clarify its scope. A natural worry is that it might legitimise censorship of ordinary users in immersive environments. While it is true that most of what people say can, at least in principle, have political implications, not all speech is made as political speech or in a political context. For political actors, whose interventions are extensions of their ideological commitments, the application of PAFE is straightforward. Things become murkier once we move beyond these cases, so applying PAFE requires recognising that political speech in immersive environments lies on a spectrum. At one end, we have explicitly political agents whose aim is to gain or maintain power and who should be held accountable for how they express themselves through their avatarbots. At the other end are individuals who exercise their free speech rights primarily to gain an audience rather than power; comedians are an obvious example. As long as they do not convert that audience into political capital, PAFE does not apply to them. In between these poles are actors such as think tanks and NGOs, who both seek audiences and aim to shape political outcomes; here, whether PAFE should apply depends on the extent to which their communication is tied to democratic decision-making.

These categories are not meant to suggest that there is a clean conceptual line be-

tween political and non-political expression in immersive environments (or in general, for that matter). In practice, almost any interaction can have political implications, and attempts to police political content at the level of individual utterances would be both unrealistic and normatively problematic. Rather than treating PAFE as a universal filter for all speech, we understand it as a standard that could apply primarily to civic spaces where communication explicitly claims political authority, to actors who occupy roles that carry democratic responsibilities (such as parties, candidates, and public officials), and to organised campaigns that are directed at shaping electoral or policy outcomes.

With this clarification in place, our next step is to situate PAFE within the existing European regulatory landscape. Section 3.1 shows how instruments such as the Digital Services Act and the AI Act partially reflect PAFE's concerns with authenticity, non-domination, and verifiability, before Section 3.2 turns to design principles for metaverse platforms.

3.1. Towards a European regulatory framework: the Digital Services Act (DSA) and the EU AI Act

Although PAFE is not a legal principle in existing European Union legislation, the concerns expressed by its three criteria are addressed to varying degrees by the EU Artificial Intelligence Act, the Digital Services Act, and Regulation (EU) 2024/900 on the transparency and targeting of political advertising. This section uses PAFE as an interpretive framework for comparing these instruments. It asks whether their provisions address concerns that are functionally similar to those captured by PAFE, rather than whether they expressly adopt PAFE or fully satisfy its normative requirements.

TABLE**2:**

Thematic
correspondence
between
PAFE
and
selected
EU
regulatory
instruments.
“PAFE-
related
elements
addressed”
indicates
that
a
provision
directly
or
indirectly
addresses
a
concern
captured
by
authenticity,
non-
domination,
or
verifiability.
“Limitations,
exceptions,
and
PAFE-
related
gaps”
includes
restrictions
on
the
scope
of
those
provisions
and
aspects
of
PAFE
that
remain
unaddressed

(Table created by AD with the assistance of NotebookLM (Google) and ChatGPT 5 (OpenAI) and verified by the authors).

ACT/REG + PAFE CR.	PAFE-RELATED ELEMENTS ADDRESSED (THEMES + ARTICLES/RECITALS)	LIMITATIONS, EXCEPTIONS, AND UNADDRESSED ELEMENTS (THEMES + ARTICLES/RECITALS)
EU AI Act Authenticity	- Watermarking / Machine-readable marking for synthetic content (Art. 50(1),(2)) - Disclosure of AI-generated deep fakes (Art. 50(4)) - Disclosure for AI-generated public-interest text (Art. 50(2)) - Ban on manipulative/subliminal AI (Art. 5(1)(a))	- Exceptions for artistic/fictional deep fakes (Art. 50(4)) - Exemption when human editorial review exists (Art. 50(4)) - Law-enforcement exceptions for disclosures/markings (Art. 50(1),(2),(4)) - Non-coverage of administrative/logistical political tools (Annex III, 8(b))
EU AI Act Non-Domination	- Ban on exploiting vulnerabilities (Art. 5(1)(b)) - Ban on social scoring (Art. 5(1)(c)) - High-risk classification for election-influencing AI (Annex III, 8(b))	- Lawful evaluation practices allowed (Recital 34) - Commercial practices allowed if lawful (Recital 32)
EU AI Act Verifiability	- Logging for high-risk AI (Art. 12(1),(2)) - Specific biometric logging (Art. 12(3)) - Technical documentation (Art. 11; Annex IV) - GPAI documentation duties (Art. 53(1)(a),(b); Annex XI, XII)	- Exclusion of sensitive operational data (Art. 26(5); Art. 72(2)) - Open-source GPAI exemption (Art. 53(1),(2)) - Online spaces excluded from restrictions ("publicly accessible space" definition)
Digital Services Act Authenticity	- Measures against fake accounts/bots (Recital 84) - Statements of reasons for content restriction (Art. 17(1),(3)(b)) - Trader traceability (Art. 30(1),(2))	- No general monitoring obligation (Art. 8; Recital 30) - Limited disclosure of notifier identity (Recital 54) - Deceptive commercial content exception (Recital 55) - Information orders limited to specific recipients (Recital 37)
Digital Services Act Non-Domination	- Identification of ad payer (Art. 26(1)(c)) - Ban on targeting using sensitive data (Art. 26(3)) - Advertising systems included in	- No general advertising prohibition (Recital 10, 68) - Recommender systems still allowed (Art. 27(1)) - Dark-pattern limits apply only to DSA scope (Recital 67)

ACT/REG + PAFE CR.	PAFE-RELATED ELEMENTS ADDRESSED (THEMES + ARTICLES/RECITALS)	LIMITATIONS, EXCEPTIONS, AND UNADDRESSED ELEMENTS (THEMES + ARTICLES/RECITALS)
	systemic-risk analysis (Recital 84, 88) • Ad repositories (Art. 39(1); Recital 95)	
Digital Services Act Verifiability	• Orders to provide information (Art. 10(1)) • EU legal representative requirement (Art. 13(1)) • Trader identity verification (Art. 30(1),(2))	• No general active fact-finding (Art. 8; Recital 30) • Data minimisation for minors (Recital 71) • Only existing data must be provided (Art. 10(2)(b))
Regulation 2024/900 Political Advertising Authenticity	• Sponsor identification (Art. 1(3); Art. 12(1)) • Detailed transparency notices (Art. 12(1)) • Disclosure of remuneration source (Art. 12(1)(b)) • Declaration and verification of political nature (Art. 7(1))	• Personal political opinions excluded (Art. 1(3)) • Editorial content excluded unless paid (Art. 1(2)) • Ancillary service providers excluded (Art. 3(6)) • Unpaid user content excluded (Art. 3(5))
Regulation 2024/900 Non-Domination	• Restrictions on third-country sponsors during electoral periods (Art. 5(2)) • Financial transparency of benefits and funds (Art. 12(1)(d),(e)) • Ban on targeting using special categories of data (Art. 18(1)(c))	• No change to existing campaign-finance rules (Art. 2(2)) • Free allocation of public spaces excluded (Art. 3(2)(iii)) • Content of ads not regulated (Art. 2(2)) • SMEs exempt from annual reporting (Art. 14(2))
Regulation 2024/900 Verifiability	• Record-keeping for 7 years (Art. 9(1),(3)) • EU repository for political ads (Art. 13(1)) • Machine-readable transparency (Art. 9(3); Art. 12(3)) • Transparency on targeting logic (Art. 19(1)(c)) • Access for vetted researchers, journalists (Art. 17(1),(2))	• Record-keeping exemption for micro-undertakings (Art. 9(4); Art. 12(5)) • Ancillary services excluded from provider definition (Art. 3(6))

Both the DSA and the EU AI Act help us in framing the impact of avatarobots in terms of risk. They provide us with key insights on risk *assessment* and risk *mitigation*. For instance, Article 34(1) of the DSA mandates that the owners of platforms should identify, analyse, and assess any systemic risks (Regulation 2022/2065). Systemic risks are explicitly understood as risks of actual/foreseeable negative effects on a wide range of democratic processes, civic/political discourse and electoral processes, in particular as they relate to the sharing of illegal content, encroachment on fundamental rights and manipulation (see Efroni 2021). In a similar vein, the EU AI ACT echoes the DSA's framing of systemic risks in Recital 110, with its explicit mention of how General-Purpose AI models could pose a threat to democratic values and human rights. The EU Act goes beyond the DSA by classifying some AI systems as being “high risk” when they are “intended to be used to influence the outcome of an election or referendum or the voting behaviour of natural persons in the exercise of their vote in elections or referenda” (Regulation (EU)

2024/1689).

In light of these risks, both the DSA and AI Act advance regulatory directions that overlap with PAFE. Systems deemed to pose unacceptable risks are to be prohibited, while those classified as high risk are subject to enhanced obligations rather than outright bans. Service providers are required to consider how the design of their services contributes to disinformation or manipulative practices, to establish crisis protocols, and to increase transparency around AI-generated content. Taken together, these instruments frame AI-driven political manipulation (especially in the context of elections) as a matter of systemic risk. In the remainder of this section, we build on this regulatory baseline and ask what else could be done to make immersive environments public deliberation spaces.

3.2. Democratic design

Whether authenticity, non-domination, and verifiability are respected in immersive environments will depend on the design and governance of these platforms and not just on users' efforts. This is why, in the case of cyborg propaganda, Kunst et al. (2026) argue for structural rather than content-based regulation: safeguarding authenticity, non-domination, and verifiability requires governing the architectures of coordination, amplification, and disclosure through which political influence is organised, rather than policing the content of individual messages. In what follows, we advance some principles and design guidelines for a metaverse that fosters a healthy democratic and civic life, rather than definitive technical or institutional solutions.

The first criterion, authenticity, requires that political expression reflect the reasons of an agent who holds them. But immersive environments can be disorienting, as they offer rich sensorial cues but impoverished contextual information. A user who encounters an avatar cannot immediately know whether it is human or synthetic, grassroots, or sponsored. A legislative framework could therefore require or encourage platforms to provide epistemic-orientation mechanisms, such as dashboards of provenance, traceable histories of coordinated and organised political avatar interaction, and public indicators of network affiliations and sponsorship. Borrowing from O'Neill's call for epistemic trustworthiness, users must be equipped not just to receive content, but to evaluate it.

But just as liberal democracies depend partly on civic education, immersive democratic spaces may benefit from forms of civic onboarding, including tutorials, shared norms, and interface cues that could help participants interpret synthetic

presence, recognise coordinated campaigns, and understand the provenance of political communication. Such measures should focus on increasing users' capacity to understand political communication and how immersive environments actually work.

When it comes to non-domination, the problem is that at present, most immersive platforms are owned by private corporations whose governance frameworks are optimised for engagement and revenue rather than public interest. This means that political communication can't be perfectly insulated from the covert conversion of economic power into political influence. But non-domination can require something a bit narrower: that the conversion of economic and computational resources into apparent civic support be blocked where it is covert, and made visible where it is not.

Another consideration for a future legislative framework is the development of forms of platform constitutionalism informed by Walzer's concern with sphere domination. Applied here, democratic institutions may need to set the basic rules of immersive public life rather than leaving them to platform operators. Possible measures could include open governance charters, representative user councils, and transparent procedures for identity verification, content governance, and policy change (Shapiro & Talmon, 2022). It may also require firewalls that limit the purchase of political visibility, caps on paid amplification, bans on digital doppelgängers, and restrictions on behavioural targeting in civic spaces. Public funding for non-commercial civic platforms could offer venues where political presence is not a function of marketing budgets.

A future legislative framework could also consider whether platforms should provide spaces specifically designed for political deliberation. These agoras must differ structurally from commercial platforms. Possible requirements might include transparent governance rules, independent auditing, representative user participation, declared institutional affiliations, and visibility into the funding of organised political interventions. Stronger identity and disclosure requirements should apply primarily to candidates, parties, public officials, campaign organisations, sponsors, and operators deploying synthetic agents, rather than to every citizen participating in political discussion.

Verifiability requires that citizens be able to trace a contribution to its source. Verifiability is the criterion that most directly conflicts with other democratic goods, such as anonymity and pseudonymity which can serve dissidents, whistleblowers, vulnerable groups, and ordinary citizens. Requiring every participant to reveal their

civil identity could create risks of surveillance, exclusion, and chilling effects on political participation. A legislative framework or governance mechanisms should instead consider proportionate provenance requirements for communication that claims institutional or civic authority or forms part of an organised attempt to influence electoral or policy decisions. Within that scope, several measures appear possible. Digital duplicates speaking on behalf of public figures might be entered in a public register of authorised political avatars, allowing users to confirm that a given avatar is what it claims to be. Non-consensual replicas would fall under existing prohibitions on impersonation and could be subject to removal and sanction. Autonomous agents engaged in political communication could be required to carry accessible metadata identifying their deployer.

Of course, these considerations are not meant to provide a definitive design for a democratic metaverse. Rather, they identify the issues that future legislation and platform governance would need to address if immersive environments are to foster accountable political expression: how users can recognise synthetic and sponsored political communication, how concentrated resources can be prevented from generating artificial political presence, and how responsibility can be made verifiable without eliminating privacy, anonymity, or meaningful political participation.

4. Conclusion

Any technological advance that touches the public sphere must be evaluated not only by its affordances, but by its compatibility with deliberative agency. This is something we did too late with social media platforms. Avatarbots, if left unchecked, risk becoming the dominant voices of our immersive future.

The age of avatarbots compels us to rethink not only the technologies we design, but the democratic principles we take for granted. Immersive environments, with their promise of realism and relationality, may appear to revive democratic dialogue. In reality, they risk transforming deliberation into simulation and participation into performance. Immersive environments governed by market logic and corporate incentives may fail to develop into democratic public spaces. Instead, they risk becoming environments in which synthetic voices crowd out accountable real citizens' political agency.

Yet the future is not fixed. The appropriate response to the dangers sketched in this paper is not technological alarmism, but careful institutional, legal, ethical, and infrastructural design. A truly democratic metaverse must embed the principle of *accountable free expression*: speech tied to identifiable, responsible agents; inter-

actions grounded in verifiable contexts; and influence traceable through transparent mechanisms.

ACKNOWLEDGEMENTS

Constantin Vica and Radu Uszkai received funding by the European Union (ERC, avataResponsibility, 101117761). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

The work of Alexandru Dancu is supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI - UEFISCDI, project number PN-IV-P8.1-PRE-HE-ORG-2024-0174, funded through the University of Bucharest under Contract No. 66PHE/2024.

Cristina Voinea's work has received funding from the project Counterfactual Assessment and Valuation for Awareness Architecture – CAVAA (European Commission, EIC 101071178).

References

- Administrația Prezidențială. (2025). *Analiză privind războiul hibrid derulat de Federația Rusă împotriva României [Analysis of the hybrid war conducted by the Russian Federation against Romania]*. Parchetul de pe lângă Înalta Curte de Casație și Justiție. <https://observatornews.ro/uploads/modules/apps/633300/R%C4%83zboi%20hibrid.pdf>
- AISE. (2025). *Relazione annuale sulla politica dell'informazione per la sicurezza 2025 [Annual report on security intelligence policy]*. Sistema di informazione per la sicurezza della Repubblica. <https://www.sicurezzanazionale.gov.it/data/cms/posts/986/attachments/aae81005-6c29-487b-85cc-b0a310f8c9be/download?view=true>
- Bolter, J. D., & Grusin, R. (2000). *Remediation: Understanding new media*. MIT Press.
- Bradshaw, S., & Howard, P. N. (2018). *Challenging truth and trust: A global inventory of organized social media manipulation* (pp. 1–25) [Project on Computational Propaganda]. Oxford Internet Institute. <https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2018/07/ct2018.pdf>
- Carpenter, J., Brady, W., Crockett, M., Weber, R., & Sinnott-Armstrong, W. (2021). Political polarization and moral outrage on social media. *Connecticut Law Review*, 52(3), 1107–1120.
- CCR. (2024). *Hotărârea nr. 32 din 6 decembrie 2024 privind anularea procesului electoral cu privire la alegerea Președintelui României din anul 2024 [Decision no. 32 of 6 December 2024 regarding the annulment of the electoral process concerning the election of the President of Romania in 2024]*. Curtea Constituțională a României. https://www.ccr.ro/wp-content/uploads/2024/12/Hotarare_32_2024.pdf

- Chalmers, D. J. (2022). *Reality+: Virtual worlds and the problems of philosophy*. W.W. Norton.
- Cheng, R., Wu, N., Chen, S., & Han, B. (2022). Will metaverse be NextG Internet? Vision, hype, and reality. *IEEE Network*, 36(5), 197–204. <https://doi.org/10.1109/MNET.117.2200055>
- Constantinescu, M. (2026). Do GenAI avatars open new responsibility gaps? *AI & Society*, 41(3), 2059–2068. <https://doi.org/10.1007/s00146-025-02660-9>
- Deibert, R. J. (2019). The road to digital unfreedom: Three painful truths about social media. *Journal of Democracy*, 30(1), 25–39. <https://doi.org/10.1353/jod.2019.0002>
- Efroni, Z. (2021, November 16). *The Digital Services Act: Risk-based regulation of online platforms*. Internet Policy Review. <https://policyreview.info/articles/news/digital-services-act-risk-based-regulation-online-platforms/1606>
- Egliston, B., & Carter, M. (2021). Critical questions for Facebook's virtual reality: Data, power, and the metaverse. *Internet Policy Review*, 10(4). <https://doi.org/10.14763/2021.4.1610>
- European Digital Media Observatory. (2025, April 29). *Analysis of social media presence of Romanian presidential candidates*. European Digital Media Observatory. <https://edmo.eu/edmo-news/analysis-of-social-media-presence-of-romanian-presidential-candidates/>
- European External Action Service. (2025). *3rd EEAS report on foreign information manipulation and interference threats exposing the architecture of FIMI operations*. European External Action Service (EEAS). <https://www.eeas.europa.eu/sites/default/files/documents/2025/EEAS-3nd-ThreatReport-March-2025-05-Digital-HD.pdf>
- European Parliament and Council of European Union. (2024b). *Regulation (EU) 2024/900 of the European Parliament and of the Council of 13 March 2024 on the transparency and targeting of political advertising*. <https://eur-lex.europa.eu/eli/reg/2024/900/oj>
- European Parliament and Council of European Union. (2024a). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- Farjam, M. (2020). The bandwagon effect in an online voting experiment with real political organizations. *International Journal of Public Opinion Research*, 33(2), 412–421. <https://doi.org/10.1093/ijpor/edaa008>
- Farjam, M., & Loxbo, K. (2024). Social conformity or attitude persistence? The bandwagon effect and the spiral of silence in a polarized context. *Journal of Elections, Public Opinion and Parties*, 34(3), 531–551. <https://doi.org/10.1080/17457289.2023.2189730>
- Floridi, L. (2024). Hypersuasion – On AI's persuasive power and how to deal with it. *Philosophy & Technology*, 37(2), article number 64. <https://doi.org/10.1007/s13347-024-00756-6>
- Funky Citizens. (2024). *Undermining democracy: The weaponization of social media in Romania's 2024 elections*. <https://funky.org/en/undermining-democracy-the-weaponization-of-social-media-in-romania-2024-elections/>
- Gattermann, K., & Tulin, M. (2025). TikTok use and young AfD voters. In M. Kaeding, M. Müller, & A. Hoppe (Eds), *The 2024 European Parliament elections: A turn to the right in the shadow of war* (pp. 319–328). Springer. https://doi.org/10.1007/978-3-031-89455-8_29

Global Witness. (2025, May 15). *TikTok algorithm continues to push multiple times more far-right content to users ahead of Romanian election*. Global Witness. <https://globalwitness.org/en/campaigns/digital-threats/tiktok-algorithm-continues-to-push-multiple-times-more-far-right-content-to-user-s-ahead-of-romanian-election/>

Human Rights Watch. (2022). *Trapped in a web: The exploitation of personal data in Hungary's 2022 elections*. Human Rights Watch. <https://www.hrw.org/report/2022/12/01/trapped-web/exploitation-personal-data-hungarys-2022-elections>

Hurshman, C. E., Voinea, C., Constantinescu, M., D'Alessandro, W., Feroz, F., Hongladarom, S., Jurcys, P., Kozlovski, A., Lange, B., Liu, P., Menikoff, J., Nyholm, S., Puzio, A., Schwitzgebel, E., Sharma, B. G., Sweeney, P., Vică, C., Weissglass, D. E., Zahiu, A., & Earp, B. D. (2026). Ethics of 'digital duplicates' or AI simulations of real people: Towards an international consensus. *JME Practical Bioethics*, 2(3), e000054. <https://doi.org/10.1136/jmepb-2025-000054>

Kunst, J. R., Bierwaczzonek, K., Cha, M., Ebrahimi, O. V., Fawcett-Atkinson, M., Følstad, A., Gollwitzer, A., Köbis, N., Marcus, G., Roozenbeek, J., Schroeder, D. T., Bavel, J. J. V., Linden, S. van der, White, R., & Wilhelmsen, L. L. (2026). *Puppets or partners? Governing cyborg propaganda in the digital public square*. arXiv. <https://doi.org/10.48550/arXiv.2602.13088>

Kyrlitsias, C., & Michael-Grigoriou, D. (2018). Asch conformity experiment using immersive virtual reality. *Computer Animation and Virtual Worlds*, 29(5), e1804. <https://doi.org/10.1002/cav.1804>

Kyrlitsias, C., Michael-Grigoriou, D., Banakou, D., & Christofi, M. (2020). Social conformity in immersive virtual environments: The impact of agents' gaze behavior. *Frontiers in Psychology*, 11, article 2254. <https://doi.org/10.3389/fpsyg.2020.02254>

Lazar, S. (2024). *Automatic authorities: Power and AI*. arXiv. <https://arxiv.org/abs/2404.05990>

Levy, N. (2022). *Bad beliefs: Why they happen to good people*. Oxford University Press. <https://doi.org/10.1093/oso/9780192895325.001.0001>

Li, K., Mostajeran, F., Rings, S., Hertel, J., Schmidt, S., Arz, M., & Steinicke, F. (2026). Anthropomorphic AI: A toolkit for authoring and interacting with intelligent virtual agents for extended reality. *Frontiers in Virtual Reality*, 7, 1794720. <https://doi.org/10.3389/frvir.2026.1794720>

Marin, L., & Vică, C. (2025). Hic sunt leones. User orientation as a design principle for emerging institutions on social media platforms. *AI & Society*, 40(3), 1613–1626. <https://doi.org/10.1007/s00146-024-01932-0>

Marocico, O., Mirodan, S., & Ings, R. (2025, September 21). How Russian-funded fake news network aims to disrupt election in Europe – BBC investigation. *BBC News*. <https://www.bbc.co.uk/news/articles/c4g5kl0n5d2o>

Matthes, J., Knoll, J., & von Sikorski, C. (2018). The "spiral of silence" revisited: A meta-analysis on the relationship between perceptions of opinion support and political opinion expression. *Communication Research*, 45(1), 3–33. <https://doi.org/10.1177/0093650217745429>

McNair, B. (2018). *An introduction to political communication* (6th edn). Routledge.

Mihailov, E., Voinea, C., & Vică, C. (2023). Is online moral outrage outrageous? Rethinking the indignation machine. *Science and Engineering Ethics*, 29(2), article number 12. <https://doi.org/10.1007/s11948-023-00435-3>

Mill, J. S. (1989). *On liberty and other writings*. Cambridge University Press.

- Mooij, A. (2025). *Synthetic manipulation and electoral integrity: Deepfakes, bots, and the limits of the EU's Democratic Defense Package* [Working paper]. SSRN. <https://doi.org/10.2139/ssrn.5825006>
- Nadeau, R., Cloutier, E., & Guay, J. H. (1993). New evidence about the existence of a bandwagon effect in the opinion formation process. *International Political Science Review*, 14(2), 203–213. <http://doi.org/10.1177/019251219301400204>
- Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme*, 17(2), 141–161. <https://doi.org/10.1017/epi.2018.32>
- Nikolaou, A., Schwabe, A., & Boomgaarden, H. (2022). Changing social attitudes with virtual reality: A systematic review and meta-analysis. *Annals of the International Communication Association*, 46(1), 30–61. <https://doi.org/10.1080/23808985.2022.2064324>
- O'Neill, O. (2020). Trust and accountability in a digital age. *Philosophy*, 95(1), 3–17. <https://doi.org/10.1017/S0031819119000457>
- Raile, E. D., Raile, A. N., Salmon, C. T., & Post, L. A. (2014). Defining public will. *Politics & Policy*, 42(1), 103–130. <https://doi.org/10.1111/polp.12063>
- Rosenberg, L. (2023). The metaverse as an AI-mediated tool of targeted persuasion. In M. Zallio (Ed.), *Human-centered metaverse and digital environments* (Vol. 99, pp. 122–130). AHFE Open Access. <https://doi.org/10.54941/ahfe1003938>
- Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., & Stieglitz, S. (2019). Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, 28(4), 394–412. <https://doi.org/10.1080/0960085X.2018.1560920>
- Schmitt-Beck, R. (2008). Bandwagon effect. In W. Donsbach (Ed.), *The international encyclopedia of communication* (Vol. 2, pp. 308–310). Wiley-Blackwell.
- Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., Goldenberg, A., Kyrychenko, Y., Leyton-Brown, K., Lutz, N., Marcus, G., Menczer, F., Pennycook, G., Rand, D. G., Schweitzer, F., Summerfield, C., Tang, A., Van Bavel, J., van der Linden, S., & Kunst, J. R. (2025). How malicious AI swarms can threaten democracy. *Science*, 391(6783), 354–357. <https://doi.org/10.1126/science.adz1697>
- Shapiro, E., & Talmon, N. (2022). *Foundations for grassroots democratic metaverse*. arXiv. <https://doi.org/10.48550/arXiv.2203.04090>
- VIGINUM. (2025). *Manipulation d'algorithmes et instrumentalisation d'influenceurs: Enseignements de l'élection présidentielle en Roumanie & risques pour la France* [Algorithm manipulation and the instrumentalization of influencers: Lessons from Romania's presidential election & risks for France]. VIGINUM. <https://www.sgdsn.gouv.fr/publications/manipulation-dalgorithmes-et-instrumentalisation-dinfluenceurs-enseignements-de>
- Voinea, C., Marin, L., & Vică, C. (2024). Digital slot machines: Social media platforms as attentional scaffolds. *Topoi*, 43(3), 685–695. <https://doi.org/10.1007/s11245-024-10031-0>
- Voinea, C., Porsdam Mann, S., Savulescu, J., & Earp, B. D. (2025). Digital doppelgängers, human relationships, and practical identity. *Bioethics*, 40(5), 435–444. <https://doi.org/10.1111/bioe.70026>
- Voinea, C., Vică, C., Mihailov, E., & Savulescu, J. (2020). The internet as cognitive enhancement. *Science and Engineering Ethics*, 26(6), 2759–2780. <https://doi.org/10.1007/s11948-020-00210-8>

Walzer, M. (1983). *Spheres of justice: A defense of pluralism and equality*. Basic Books.

Whittaker, L., Weismueller, J., & Pallant, J. I. (2026). Avatars, assemble! Digital human avatars for public figure engagement and marketing. *Business Horizons*, 69(4), 449–461. <https://doi.org/10.1016/j.bushor.2025.06.004>

Woolley, S. (2023). *Manufacturing consensus: Understanding propaganda in the era of automation and anonymity*. Yale University Press.

Woolley, S. C., & Guilbeault, D. R. (2017). *Computational propaganda in the United States of America: Manufacturing consensus online* (pp. 1–28) [Computational Propaganda Research Project]. Oxford Internet Institute. <https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2017/06/Comprop-US A.pdf>

Woolley, S. C., & Howard, P. N. (Eds). (2018). *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford University Press.

Zuiderveen Borgesius, F. J., Möller, J., Kruikemeier, S., Ó Fathaigh, R., Irion, K., Dobber, T., Bodo, B., & de Vreese, C. (2018). Online political microtargeting: Promises and threats for democracy. *Utrecht Law Review*, 14(1), 82–96. <https://doi.org/10.18352/ulr.420>

Published by



ALEXANDER VON HUMBOLDT
INSTITUTE FOR INTERNET
AND SOCIETY



RESEARCH
FOR THE
DIGITAL AGE

in cooperation with



CREATE



centre
— internet
et **societe**



R&I IN3
Internet
interdisciplinary
Institute
Universitat Oberta de Catalunya



UNIVERSITY OF TARTU
Johan Skytte Institute of
Political Studies