International Journal of Communication 19(2025), 3736–3760 1932–8036/20250005

Synthetic Diversity: Examining the Effects of Ethnic Targeting Using AI-Generated Political Ads

MORGAN WACK
University of Zurich, Switzerland

DOUGLAS A. PARRY
Vrije Universiteit Amsterdam, The Netherlands
Stellenbosch University, South Africa

Can deceptive use of generative artificial intelligence (AI) in advertisements influence support for political parties? Drawing on a survey experiment based on the tactics of a secessionist movement in South Africa, this study assesses whether targeting out-group ethnicities with diverse deepfake avatars can alter voter support. We iterate from existing work by varying the ethnicity of the presenter to assess whether AI-assisted alterations to author ethnicities can mobilize support for minority parties. The results indicate that coethnic speakers are more effective at mobilizing both personal and perceptions of societal support, even when content includes AI labels. Additional analyses indicate that AI literacy scores better predict respondents' abilities to identify synthetic content than digital literacy and that coethnic avatars reduce skepticism toward AI-assisted messaging campaigns. The results of the study have implications for researchers and policy makers interested in understanding how synthetic media may be used for deceptive purposes in political campaigns.

Keywords: disinformation, political advertising, artificial intelligence, elections, deepfakes, synthetic media

Popstar Taylor Swift's endorsement of presidential candidate Donald Trump ahead of the 2024 U.S. election was widely publicized and reposted online by the candidate's account. Although the support of one of the world's most famous musicians brought attention to the campaign, there was one glaring issue with the endorsement: it never happened (Looker, 2024). Along with the publication of images purporting to show the Republican nominee surrounded by African American supporters, Taylor Swift's likeness had been generated and modified using *generative artificial intelligence* (AI) tools (Zhang, Zhang, Zhang, & Kweon, 2023). Reminiscent of the sordid origins of deepfakes as a method for exploiting the images of female celebrities (Spivak, 2018), the growing integration of synthetic media into political spaces necessitates proactive research into its potential influences (Dan et al., 2021).

Morgan Wack: m.wack@ikmz.uzh.ch
Douglas A. Parry: d.a.parry@vu.nl
Date submitted: 2025-02-20

Despite growing attention to the use of generative AI in political domains (e.g., Simchon, Edwards, & Lewandowsky, 2024; Vaccari & Chadwick, 2020), most studies have focused on the potential reputational damage—both to individual candidates and to wider democratic institutions—that these technologies can produce (Vaccari & Chadwick, 2020). Deepfakes, or synthetic images and videos of real people, have come to be associated with their use by belligerent actors aiming to cause harm (Botha & Pieterse, 2020; Langa, 2021). Although this work has helped develop policies that aim to defend democracies against the potential consequences of AI-based attacks and their use in disinformation campaigns (Chesney & Citron, 2019), many of the identified instances of synthetic media use in politics have involved direct targeting of opposition figures. Rather, as in the case of the falsified Taylor Swift endorsement, several cases have seen synthetic tools used to create deceptive materials that are aimed at generating or simulating support for candidates and political parties (LaChapelle & Tucker, 2023). To distinguish this form of deployment from the oppositional connotations of "deepfakes," recent work has described these proattitudinal use cases to be part of a new class of "softfakes" that involve "images, videos or audio clips that are doctored to make a political candidate seem more appealing" (Chowdhury, 2024, p. 237).

The most noticeable use of synthetic media to promote political campaigns has been in campaign advertisements. Despite being banned in several countries, the use of AI in the development of political ads has sparked controversies in countries as distant as Canada (Mackin, 2024), India (Shukla & Schneier, 2024), and Kenya (Kigwiru, 2022). More recently, research on the influence of advertisements generated using AI tools to target political audiences found them to be more persuasive than nonpersonalized forms of advertising (Simchon et al., 2024). Building on this work and the potential for AI tools to enable targeted advertising beyond politics (Haleem, Javaid, Qadri, Singh, & Suman, 2022), we draw on recent real-world use cases to ask: Does deceptive use of generative AI tools enable political parties to effectively target out-group voters?

To investigate this question, we conducted a preregistered, survey-based experiment using a politically relevant and locally grounded case: the Cape Independence movement in South Africa. Although the movement has limited electoral support, it has generated significant media attention because of its controversial calls for secession, its association with White minority interests, and its attempts to rebrand as a more inclusive, multicultural campaign. In response to its marginal status and image challenges, groups aligned with the movement have turned to generative AI tools to create synthetic campaign materials that depict non-White supporters. This context allowed us to investigate not only the persuasive potential of generative AI in targeting politically and ethnically distant voters but also how detection of synthetic content and viewer identity might moderate these effects. Finally, as movements are normalized not only by shifting opinions of their audiences but also by altering their perceived support among the wider population, we consider the influence of deepfakes on both personal and societal perceptions.

Synthetic Media in Politics

The rapid improvement in the realism of synthetic media has prompted research into their use in perpetuating political harms, including the creation of false evidence of corruption (Gregory, 2022),

illusory political statements (Scott, 2024; Seitz-Wald, 2024), and blackmail (de Rancourt-Raymond & Smaili, 2023). Disinformation in the form of synthetic audio and deepfake videos purporting to catch politicians engaging in corruption or impropriety has become relatively common during campaigns. In one infamous case, the release of a faked audio clip of a Slovakian presidential candidate on the eve of the country's election has been widely cited as having contributed to his loss at the polls (de Nadal & Jančárik, 2024). In experimental settings, research has detailed how deepfakes can be used to reduce support for candidates and political parties (Dobber, Metoui, Trilling, Helberger, & De Vreese, 2021), as well as how citizens already struggle to distinguish between real and AI-generated political text (Kreps, McCain, & Brundage, 2022).

At the same time, researchers have emphasized the potential for generative content to exacerbate indirect harms on society. Should the technology advance to the point at which deepfake videos can be created to give the appearance of any individual doing any number of illicit or compromising activities without the tools to discern veracity, as many argue is already possible (Cooke, Edwards, Barkoff, & Kelly, 2024; Frank et al., 2024; Pocol, Istead, Siu, Mokhtari, & Kodeiri, 2023), there is valid concern that we will begin to distrust true content as a result. Although skepticism can be healthy and even the preferred mode of inquiry in certain domains, perpetuating distrust can reduce trust in critical institutions such as the media (Vaccari & Chadwick, 2020) and erode broader social cohesion (Matthews & Kidd, 2023; Ternovski, Kalla, & Aronow, 2022). Moreover, in the political arena, the growing ubiquity of deepfakes has already seen authorities attempt to discredit real incidents by casting doubt on the veracity of multimedia content (Christopher, 2023). Often described as the "liar's dividend," there is growing concern that as the prevalence of inauthentic content rises, real instances of impropriety from politicians are likely to become less reputationally damaging (Chesney & Citron, 2019). In support of these fears, recent experimental work has illustrated how the presence of inauthentic content increases the value of lying for politicians with relation to alternatives such as apologizing or avoiding the topic (Schiff, Schiff, & Bueno, 2025).

Beyond Reputational Harm

Despite concerns about use cases related to reputational harm and impersonation, generative content is not, and neither has it always been, used solely to harm political candidates or campaigns. Just as propaganda videos have been created to both dehumanize populations and lionize copartisans (Hobbs, 2020), generative content has also been used to both disparage (Sharma, 2024) and promote political actors and their actions (Martin, Jackson, Trauthig, & Woolley, 2024).

A recent overview of the literature found the study of deepfakes to be overwhelmingly focused on cases from the United States and Europe (Birrer & Just, 2024), despite these regions having far better and more readily available traditional media alternatives, while also remaining the best equipped to investigate the veracity of synthetic content. The limited evidence that is available from contexts beyond the Global North suggests that the use of deepfakes is more widely accepted in the political sphere by both candidates themselves and the public. For instance, in India, generative AI tools were used to enable the translation of candidate messages to reach typically excluded populations as well as to answer questions about government benefits (Shukla & Schneier, 2024). Similarly, in South Korea, politicians have found public

support for their efforts to facilitate greater interactions with the public through the creation of artificial "avatars," or digitized icons or figures (see Figure 1) that allow people to ask virtual candidates about their policy positions (Henrickson, 2023).

Between these two extremes are efforts to promote parties, candidates, and policies that use synthetic media either deceptively or in a manner beyond the scope of accepted political practice. The practice that has received the greatest attention in this space has been the use of AI to aid in "microtargeting" political advertisements. In this context, microtargeting refers to the process of directly targeting individuals using their data to tailor ads to match their known preferences and personal characteristics (Papakyriakopoulos, Hegelich, Shahrezaye, & Serrano, 2018). Although legality varies by jurisdiction (Bayer, 2024), the use of generative AI to both create and alter advertisements to appeal to specific voting cohorts has faced both public (Bernard, 2024) and legal opposition (Paget, 2024). These concerns are not without reason, as recent work has detailed how large public data-collection processes can be used to infer outputs relevant to political targeting (Hinds & Joinson, 2018) and how generative technologies could be used with relatively limited resource use to both target susceptible voter populations (Foos, 2024) as well as produce persuasive political messaging campaigns (Hackenburg & Margetts, 2024; Simchon et al., 2024).

In addition to tailoring messages to different audiences, synthetic media can also be used to tailor *messengers*. Politicians and activists may employ AI-generated avatars that reflect the race, gender, or age of particular constituencies, not only to signal inclusion but also to foster a greater sense of perceived social consensus around a cause. Critically, the projection of social support has been shown to have the potential to shift attitudes toward otherwise objectionable political movements (Portelinha & Elcheroth, 2016). This approach builds on long-standing findings from political psychology and communication research that draw on social identity theory, which emphasizes how identification with media characters can influence interpretation: individuals are more likely to support a movement when they perceive that others like them are involved and when they view that movement as broadly supported by society (Appiah & Liu, 2009; Cialdini, Reno, & Kallgren, 1990; Kalla & Broockman, 2020). In racially stratified societies, the public visibility of marginalized identities in political messaging can reduce skepticism, increase perceived legitimacy, and shape how messages are interpreted—particularly among audiences who rarely see themselves reflected in certain political narratives. Representation alone may signal wider acceptance and shared norms, boosting both societal and personal support. But support may also be amplified when viewers perceive a shared racial or ethnic identity with available content—an effect documented in studies of political persuasion, peer modeling, and social influence (Knobloch-Westerwick, Mothes, & Polavin, 2020). Given the dynamics inherent to the case context, we predict that:

H1: The use of non-White avatars has the potential to increase both perceived societal support and personal support for the movement.

H2: Coethnic, or "shared" identities between avatars and participants similarly increase both societal and personal support.

Countering Potential Harms of Synthetic Media

The Protective Role of AI Labels

Though synthetic media may be used deceptively to enhance the appeal of a candidate or movement among out-group members or voters, recent efforts have sought to introduce protective mechanisms that allow audiences to better detect and resist manipulation (Alon, Rahimi, & Tahar, 2024; Stuurman & Lachaud, 2022). One such intervention is the use of AI-generated content labels. These disclosures indicate that an image, video, or audio clip has been produced or altered using AI, and for video or image-based media, they typically involve discrete text labels (e.g., Dunton, 2023).

However, there is only limited evidence on the effectiveness of such labels. Some studies suggest that disclosures can alert viewers to the synthetic nature of media without necessarily reducing belief or persuasion (Clark & Lewandowsky, 2024), whereas other work acknowledges the challenges involved in determining the effectiveness of labels for AI-generated content (Wittenberg, Epstein, Berinsky, & Rand, 2025). Furthermore, most existing work on labeling has examined deepfakes in nonpolitical contexts or in relation to oppositional disinformation, leaving open questions about whether such protective mechanisms are equally effective when synthetic media is used in proattitudinal ways.

RQ1: Does labeling political advertisements as AI-generated reduce their persuasive power, particularly when those ads are targeted at out-group audiences?

The Protective Role of Digital and AI Literacy

In response to the potential negative implications of generative AI in political contexts, researchers have begun to investigate strategies to help individuals protect themselves from these effects. Much of this work has focused on the effectiveness of targeted media or news literacy interventions, which have generally been found to improve critical engagement with information and reduce susceptibility to mis- and disinformation (Lu et al., 2024). More broadly, digital skills are increasingly acknowledged as essential for navigating the information environment, particularly in the face of the challenges posed by disinformation and generative AI in democratic processes (de Vries, Piotrowski, & de Vreese, 2024; Huang, Jia, & Yu, 2024). These general digital skills include a range of competencies, such as accessing, evaluating, and effectively using digital information, as well as the technical proficiency needed to interact with online platforms (Helsper, Scheider, Van Deursen, & Van Laar, 2020; Van Deursen, Helsper, & Eynon, 2016).

Building on this foundation, algorithmic and AI-specific skills have been proposed to address the unique challenges posed by algorithms and generative AI (Long & Magerko, 2020). These skills emphasize understanding how AI technologies operate, including the mechanisms behind generative AI, algorithmic content tailoring, and strategies for recognizing AI-generated or manipulated content and the influence of algorithms on information feeds (de Vries et al., 2024; Dogruel, Masur, & Joeckel, 2022; Ng, Leung, Chu, & Qiao, 2021).

Research on news and media literacy and digital and AI skills highlights their protective role against misinformation, disinformation, and AI-generated manipulations (Lu et al., 2024), with studies indicating that individuals with higher levels of media or digital literacy are better equipped to critically evaluate misleading information, making them less susceptible to false claims. Similarly, emerging research on AI literacy suggests its potential to help individuals detect and debunk deepfakes and other AI-generated content (Long & Magerko, 2020; Ng et al., 2021). For instance, interventions that train participants to identify telltale signs of manipulated media, such as inconsistencies in facial expressions or mismatches in audiovisual synchronization, have shown promising results in reducing effects (see Huang et al., 2024, for a meta-analysis). However, questions remain about whether general digital skills alone suffice or if more targeted algorithmic or AI-specific skills are necessary to effectively detect generative content. Although some research has emphasized the value of specific interventions (e.g., Huang et al., 2024), there remains a need to investigate whether broader digital competencies or specific AI skills and knowledge are independently effective in protecting against synthetic media. This gap is particularly pronounced in developing countries, where research on these issues remains sparse despite the growing relevance of generative AI in these regions.

RQ2a: Does participants' digital competency influence their ability to detect that the video was generated with AI?

RQ2b: Does detection vary between general digital skills and those specifically related to identifying AI-manipulated content online?

Alongside questions about the efficacy of particular competencies, given the potential role of shared social identity in shaping message perception and trust (Knobloch-Westerwick et al., 2020), we also investigated whether detection of AI-generated political content varies based on whether the viewer shares the same ethnicity as the avatar shown in the video.

RQ2c: Does detection of AI-generated content differ based on shared ethnicity with the avatar?

Current Investigation

In this study, we contribute to research on the integration of deceptive uses of generative AI and synthetic media into political advertising. Though concern has grown about the potential use of these innovations to impersonate politicians (Barari, Munger, & Lucas, 2021) and target voter subsets (Dobber et al., 2021), we have also observed how political entrepreneurs have begun to adapt these strategies to target out-group voters. In this study, we examine the use of avatars in targeting ethnic voters outside of a political party's primary demographic supporter base. Specifically, we look at the development of "ethnic" avatars, or digital spokespersons given characteristics associated with specific ethnic groups. When the ethnicity of the avatar matches the ethnicity of the participant (or voter in the broader context), we consider this a coethnic avatar with a shared presenting identity. The use of AI-enhanced avatars, deepfakes, and related synthetic visual extensions poses questions about the potential for these tools to influence electoral outcomes. In this study, we aim to assess the efficacy of these efforts in the context of an ethnically oriented

secessionist campaign in the Global South. In doing so, we hope to elucidate the broader challenges that synthetic media pose for the conduct of democratic elections in the age of AI.

Methods

To address our research question and determine whether targeting out-group ethnicities with deepfake avatars can influence perceptions of both societal and personal support, we conducted a survey experiment focused on the campaign tactics of the secessionist "Cape Independence" movement in South Africa. Specifically, we used AI-generated campaign advertisements to vary the ethnicity of proindependence advocates to assess whether these new forms of media manipulation can be used to project support for minority ethnic parties. To contextualize the study, we begin with a brief overview of the Cape Independence movement, followed by a presentation of our procedure, instruments, and analysis plan. Before data collection, we preregistered our hypotheses and our research design.[1] The online supplementary materials are available via the Open Source Framework (OSF) at: https://osf.io/dy3ph/.

Case: Cape Independence

The end of apartheid in South Africa saw Nelson Mandela's African National Congress (ANC) win majority control in the country's founding democratic election in 1994 at all levels of the federated political system. Although it took until 2024 for the ANC to drop below the 50% threshold required to govern absent coalition support, the ANC's influence has varied at lower levels of government. For several successive elections, the Western Cape province has been governed by the Democratic Alliance (DA), the country's largest opposition party and the party most widely associated with the country's White minority population (Maphaka, 2021). Locally, debate has continued between residents of the "rest" of South Africa and residents of the Western Cape about the drivers of the Cape's relative prosperity, with explanations ranging from the DA's tenure and emphasis on public services to notes about the environmental endowments available in the region, alongside explanations that evoke the persistent divisions of apartheid that proscribe inequalities based on ethnic and racial stereotypes.

Perceptions of the Western Cape's differential success in recent years have galvanized several segments of the province's population to capitalize on the attention. For example, a small but, at times, vocal segment of this population has begun to champion the cause of Western Cape independence. Despite not having a centralized structure, the actions of the nebulous "Cape Independence" movement's main subsidiaries, including the *Cape Independence Party*, the *Referendum Party*, the *Cape Independence Advocacy Group*, *CapeXit*, and *Gatvol Capetonian*, have been widely tracked by the South African press (Davis, 2024). Although most supporters and the formalized groups that have taken on the cause of Cape Independence highlight both governmental and cultural "distinctiveness" that separates them from the rest of the country, reporters have rightfully pointed to the rhetoric used by the movement and its White supporter base to draw parallels to the colonial incursions that serve as the basis for many promoting the Cape's unique founding history (Sule, 2023).

[1] See https://aspredicted.org/3XP_48X.

Although the recent activities of groups associated with the movement have involved both online and offline political campaign efforts, the movement is largely perceived as a racially motivated novelty rather than a serious threat to South African sovereignty. Although surveys conducted by movement representatives claim to show support close to most provincial residents, these numbers have not been independently verified. Moreover, the South African Election Commission's (IEC) decision to require physical rather than electronic signatures to qualify to compete in the 2024 national elections resulted in a collapse in verifiable support (Charles, 2024). Amid these setbacks, the movement has aimed to improve its image among the non-White South African residents who would need to be swayed to vote in support of independence for the movement to have any real influence. This pivot has included outreach efforts along with the development of proindependence materials that stress the presence of diverse sets of supporters. At the same time, the movement's marginal influence, financial constraints, and negative media publicity pose a challenge to the recruitment of out-group (non-White) supporters. It is in this context that several of the groups that collectively contribute to the advancement of the wider Cape Independence movement have begun to use generative AI to create advertisements and campaign materials to "create" non-White supporters and advocates to appeal to South Africa's majority Black voter base.

Participants and Procedure

To assess the influence of the ethnically diverse synthetic advertisements, we recruited a sample of 1,025 participants from South Africa's Western Cape Province through Prolific. Because of difficulties balancing participants by ethnicity, data collection was divided into several identical surveys that included screening questions to reach a more diverse participant pool. These surveys, which were approved by Clemson University's Office of Research Compliance (Study ID: IRB2024-0296),[2] were activated between late May and mid-July 2024.[3] This process enabled us to obtain the participation of a diverse sample set, which was facilitated using Prolific participant quotas.[4] Participants were paid on average $2.01 for their participation in the survey, which corresponded to a rate of $10.37 per hour.[5] In addition, participants were informed that they would be viewing political advertisements.

Following data collection, we eliminated participants who completed the survey faster than was judged to have been possible given the inclusion of videos (less than five minutes) or greater than two standard deviations above the mean completion time (more than 25 minutes). We additionally excluded

[2] While exempted because of the lack of deviation from the existing advertisements, we provided participants with details about the videos that were disclosed as advertisements.

[3] We originally intended to conduct the survey before South Africa's 2024 general election to limit bias in perceptions of societal support that could occur through the publication of voting totals. Challenges with the vetting process used to identify eligible participants left us with less than half of our target sample in the preelection period. Following the IEC's removal of the Cape Independence Party from the ballot, our concern was less relevant, and we opted to continue data collection in the postelection period.

[4] Participants were excluded from retaking subsequent surveys based on their Prolific IDs to ensure the sample contained only new participants.

[5] This reimbursement amount was roughly 30% per hour more than the South African minimum wage of R27.58 ($1.55 at current exchange rates).

participants who failed an attention check as well as a small set of participants who noted issues with their audio or video that limited their ability to consider the treatment. The final data set contained 875 participants with an average completion time just under 13 minutes (769 seconds). The ethnic makeup of the sample included 151 White participants, 283 participants who are people of color,[6] and 441 Black participants, with a mean age of 30.98 (*SD* = 10.01). Table 1 presents summary statistics for the main variables in the data set.

Table 1. Summary Statistics.

Variable	N	Mean/%	Std. Dev.	Min	Max
Avatar Ethnicity					
... *White*	244	28			
... *Black*	248	28			
... *People of Color*	253	29			
... *Control*	130	15			
Respondent Ethnicity					
... *White*	151	17			
... *Black*	441	50			
... *People of Color*	283	32			
Duration (*Seconds*)		767	249	325	1492
Societal Support		49	29%	0%	100%
Personal Support		43	34%	0%	100%

Note. Control refers to respondents who were not presented with any treatment content.

The experiment followed a between-subjects design with each participant randomly assigned to one of three treatment groups presented with a video containing a political advertisement. The control group was not presented with a video and was instructed to continue on to answer questions about party support, as were the other participants.[7]

Instruments

A complete version of the survey is available via the online materials.[8]

Treatment Stimuli

Our treatment stimuli were modeled on videos released by a prominent pro-independence advocacy group in advance of South Africa's 2024 general election. The videos used AI tools to generate avatars that

[6] An official ethnic categorization in South Africa, the nation's community of people of color has an identity and history distinct from the White and Black communities. See Posel (2001) for an overview of the community's history and culture.

[7] The appendix contains further information on differences between treatment stimuli across groups in Table A.6.

[8] See: https://osf.io/dy3ph/

provided potential voters with details about the Cape Independence movement. We used *DeepReel*, a publicly accessible avatar creation company reliant on synthetic media, to re-create one of the advertisements. Specifically, we created six avatars varied by race and gender. A preliminary norming process conducted among university students in South Africa and the United States was used to evaluate content verisimilitude to ensure that the realism of the selected video stimuli was balanced across ethnicities (n = 72). This process led us to use the male avatars for the treatment because of balance across ethnicities and concern that mixed-gender videos may bias the treatment. Each video saw the avatar repeat the same advertising text,[9] which is included in the OSF repository, over a blue backdrop, as depicted in Figure 1.[10]

(a) *Avatar Template.* ***(b)*** *Avatar Template with Disclaimer.*

Figure 1. Example treatment videos.

Building on work that has illustrated the ineffectiveness of AI labels in studies of deepfakes related to fabricated criminal activities (Clark & Lewandowsky, 2024) and body image perceptions on social media (Bennett, 2024), we integrated labels to assess a commonly pitched solution to the influence of generative advertisements in political campaigns. In half of the treatment videos, a label indicating that the video used AI was printed on the right side of the screen. The note, which matched current best practices in labeling used by several online art curators (Dunton, 2023), stated: "Note: This video was created using AI tools." Before watching a video, participants were presented with the following instructions: "Please turn up your volume and watch the following clip in full prior to proceeding to the questions."

Party Support Measures

Following the video, participants were presented with questions that tested our primary dependent variables.[11] We elected to use versions of the common feelings thermometer to assess both personal and societal support for two primary reasons. The first reason related to the context of the election, which saw the independence parties barred from competing, complicate questions about voting behavior. Second, we elected to use this measure to enable comparisons across parties, as it has been used previously across

[9] We opted to mirror the language of active advertisements in part to limit the potential for novel deception.
[10] A logo linking the video to an organization promoting the independence cause was included in the original video stimuli but has been removed here for privacy reasons.
[11] These questions were randomly presented alongside questions about other political parties to make the emphasis on the Cape Independence movement less central to the reader.

both first-past-the-post and multiparty electoral systems.[12] To operationalize the societal perceptions component, we asked: "If we asked 100 residents from the Western Cape Province, how many do you think would vote in support of the Cape Independence movement?" Responses were recorded on a sliding scale ranging from 0 to 100. To further gauge perceptions of wider support, participants were asked: "If we asked 100 residents from the Western Cape Province, how many do you think would support holding a referendum on Western Cape Independence?"[13] These questions draw on prior surveys that have attempted to probe for differences between personal and societal perceptions (Chia, 2014). Finally, to judge changes in personal support, participants were asked: "How warmly do you feel about the Cape Independence Movement/Referendum Party? (0 = cold; 100 = warm)." Although not a direct measure of electoral support, we rely on this version of the feelings thermometer because of the multiparty structure of South African politics and an emphasis on affective changes in opinion.

Digital Competency Measures

At the end of the survey, we asked participants how likely they were to believe that the video was created using generative AI, with response options ranging from 0 ("very unlikely") to 4 ("very likely"). Among the 745 participants who viewed the video and answered this question, the mean was 2.84 (*SD* = 1.03).

We assessed two aspects of a participant's digital competency: *traditional digital skills* and *AI-related digital skills*. To evaluate traditional digital skills, we used the six-item *information navigation and processing* subscale from the *DSI digital skills survey* (Helsper et al., 2020). This subscale, which is an extension of the widely recognized *Internet Skills Scale* (Van Deursen et al., 2016), defines information navigation and processing as "the ability to find, select, and critically evaluate digital sources of information" (p. 15). Participants rated six statements (e.g., "I know how to check if the information I find online is true") on a 5-point Likert scale, ranging from 1 ("Not at all true of me") to 5 ("Very true of me"), with "I do not understand what you mean by this" coded as 0 and "I do not want to answer" excluded from analysis. The total score, averaged from 1 to 5, reflects the participant's level of digital skill, with higher scores indicating greater proficiency. In this sample, the mean was 4.41 (*SD* = 0.54).

To assess AI-related digital skills, we employed the *AI skills* subscale from the *DigIQ* scale (de Vries et al., 2024). Although this subscale does not specifically address generative AI, it offers insights into participants' general abilities to navigate the digital information environment. It measures two key aspects: (1) participants' knowledge of how AI systems can manipulate online content and (2) their ability to identify and exert control over how these systems present content to them. This subscale, therefore, evaluates participants' general familiarity and competence with AI-driven systems. The subscale includes six items

[12] See Gidron, Scheffer, and Mor (2022) for a longer discussion on the validity of the thermometer measurement outside the United States.

[13] We have removed this question from the study's main plots and instead opted to rely on the direct question about support for Cape Independence. In examining the results, we found limited variation from the set scale point and suspect that our failure to define the concept of a referendum (which are not common in South Africa) likely influenced the results.

(e.g., "I recognize when a website or app uses AI to adjust the content for me"), with responses on a 5-point Likert scale ranging from 1 ("Completely untrue") to 5 ("Completely true"). The average total score, which ranges from 1 to 5, indicates the participant's level of competence with AI-related content, with higher scores reflecting greater proficiency. In this sample, the mean score was 3.55 (*SD* = 0.93).

Analytical Approach

All data processing and analyses were conducted using R, and the scripts are available via the OSF repository. We transformed several variables for interpretability, as each primary dependent variable had a numeric output ranging from 0 to 100. In addition, we code the pairings between participants and treatment groups to focus on instances in which the ethnicity of the avatar matched the ethnicity of the participant. For instance, for "shared identity" to be coded a "1," Black South Africans would need to have been assigned to the treatment group containing the Black avatar. Similarly, White participants would need to have been assigned the White avatar, and participants who were people of color would need to have been assigned the avatar indicating people of color. Any other pairing would be coded "0."

We also subset our data to focus on specific participant groups for several of our analyses. For our analysis related to the ethnicity of the avatars on the treatments, we focus specifically on differences between White and Black participants, excluding participants who were people of color. We remove this subgroup as their responses are challenging to interpret, given their ties to each of the other communities. For the main shared identity analyses, we use the full sample minus the control group, which is removed as participants in the control group were not shown a video and thus do not have a treatment to match with participant ethnicity. We use the splits in labeled and unlabeled video advertisements to assess whether the labels moderate support for Cape Independence among participants who share an ethnicity with the treatment avatar. For the AI-detection analyses, we removed participants in the control condition, as they did not watch the video or receive the relevant AI-perception item. Models use standard errors clustered by respondent. We also report standardized regression coefficients (β), *p*-values, cluster-robust standard errors (SE), and *t*-values.

Findings

Synthetic Ads and Identity

Baseline support by ethnicity reflects wider support for Cape Independence. Participants in the control group, who were not exposed to the treatment, were asked to indicate their level of warmth toward Cape Independence. As expected, support was highest among White participants at 65%, followed by People of Color participants at 42%, and Black participants at 31%.

We first assess the influence of the avatar's presented ethnicity on our dependent variables directly. In the direction of our hypotheses, which expected the presentation of out-group support to alter perceptions of the movement's wider palatability, there were positive effects on both perceived societal and personal support for Cape Independence when the avatar was portrayed as a Black South African. These effects, however, do not reach statistical significance at our threshold of $p = .05$. Figure 2 displays

the results for both ethnicities. Compared to the White avatar, the Black avatar treatment shows directional effects for perceived societal support (β = 0.07, SE = 0.045, t(490) = 1.64, p = .10) and personal support (β = 0.08, SE = 0.045, t(490) = 1.79, p = .07), though these effects remain marginal and are not statistically significant.

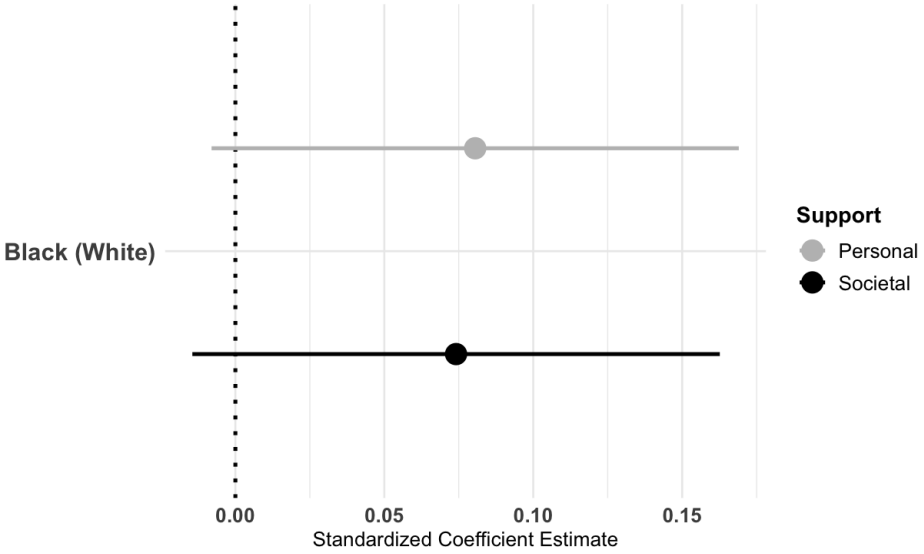

Figure 2. Influence of avatar ethnicity on personal and perceived societal support for Western Cape independence.

Note. Lines represent the 95% confidence interval. The dashed line represents the exact null effect and corresponds with Table A.1 in the Supplementary Materials.

Next, we focus on instances in which the ethnicity of the avatar matches the ethnicity of the participant (shared identity). The results of the analyses are presented in Figure 3. Among the full sample, we find marginal directional effects that do not reach significance. Specifically, we find a small but insignificant effect for perceived societal support (β = 0.04, SE = 0.037, t(743) = 1.11, p = .268) and for personal support (β = 0.03, SE = 0.036, t(743) = 0.78, p = .437) for Cape Independence. However, consistent with our hypotheses that coethnic ads could be used to target "out-group" voters by presenting the movement as one supported by in-group members, we observer larger effects when we subset the analysis to focus on the targeted Black South African cohort. Among Black participants assigned a Black avatar, the effect size for societal support becomes significant (β = 0.12, SE = 0.052, t(361) = 2.21, p = .028), and the effect on personal support doubles, though it is still not statistically significant (β = 0.1, SE = 0.055, t(361) = 1.85, p = .065).[14]

[14] Although not the focus of our study, which is on the influence of generative ads in reaching "out-groups," we include the effects among the other groups in the Appendix.

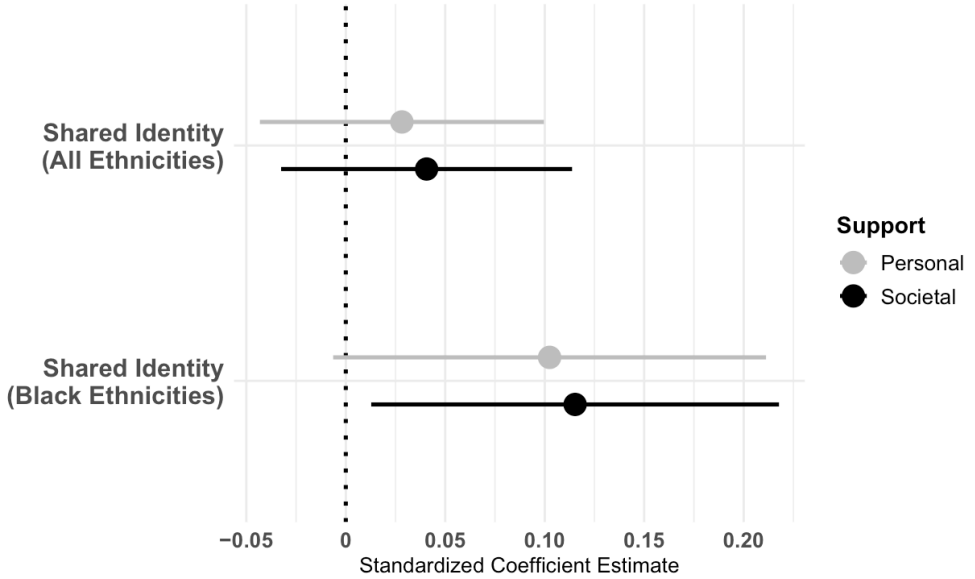

Figure 3. Influence of shared ethnicity between respondents and avatars on personal and perceived societal support for Western Cape independence.

Note. Lines represent the 95% confidence interval. The dashed line represents the exact null effect and corresponds with Tables A.2 and A.3 in the Supplementary Materials.

AI Labeling and Political Ads

To assess the influence of AI labels on the treatment advertisements, we included an interaction term with the treatment model. The findings, presented in Figure 4, both complicate and support prior work. No significant relationships were identified between the use of the AI label on either perceived societal support (β = 0.02, SE = 0.037, $t(741)$ = 0.41, p = .685) or personal support (β = −0.04, SE = 0.036, t(741) = −1, p = .318) for Cape Independence. Similarly, when we subset to focus solely on the targeted Black South African cohort, we find similar null results for perceived societal support (β = −0.01, SE = 0.046, $t(437)$ = −0.22, p = .828) and personal support (β = −0.04, SE = 0.048, $t(437)$ = −0.88, p = .382).

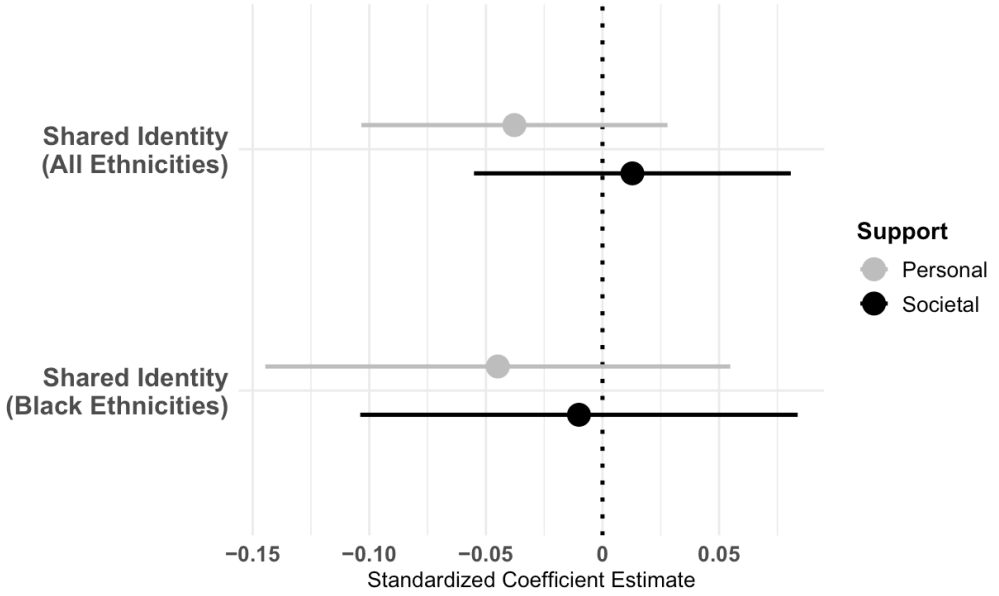

Figure 4. Influence of AI labels on personal and perceived societal support for Western Cape independence among participants with a shared identity with the avatar.

Note. Lines represent the 95% confidence interval. The dashed line represents the exact null effect and corresponds with Table A.4 in the Supplementary Materials.

Role of Digital Skills and Identity in AI Recognition

In an exploratory analysis, we investigated whether participants' ages, shared ethnic identities with the avatar in the video, and digital competencies affected their likelihood of believing that the video was created using AI. The analysis revealed no significant relationship between participants' ages ($r = -.01$, $\rho = .78$) or shared ethnicities with the avatar ($t(384.02) = 0.22$, $d = 0.02$, $\rho = .826$), and their perceptions of AI involvement. However, significant correlations were found for both digital skills ($r = .08$, $\rho = .040$) and AI skills ($r = .18$, $\rho < .001$). AI skills ($\beta = 0.18$, $\rho < .001$) remained a significant predictor of the likelihood of believing that the video was created using AI, even when controlling for more general digital skills (see Model 1 in Figure 5).

Next, we considered whether participants' shared ethnic identities with the avatar separately interacted with their general digital literacy skills (Model 2 in Figure 5) or their AI skills (Model 3 in Figure 5) to enhance or degrade the extent to which they believed that the video was produced using AI. Across these analyses, the most substantial effects were observed for AI skills ($\beta = 0.24$, $p < .001$) and the interaction between shared ethnic identity and AI skills ($\beta = -0.18$, $p < .020$) in Model 3. Here, AI skills positively predicted participants' perceptions of AI involvement in the videos, whereas conversely, the interaction between shared ethnic identity and AI skills exhibited a negative relation with their perceptions of AI involvement.

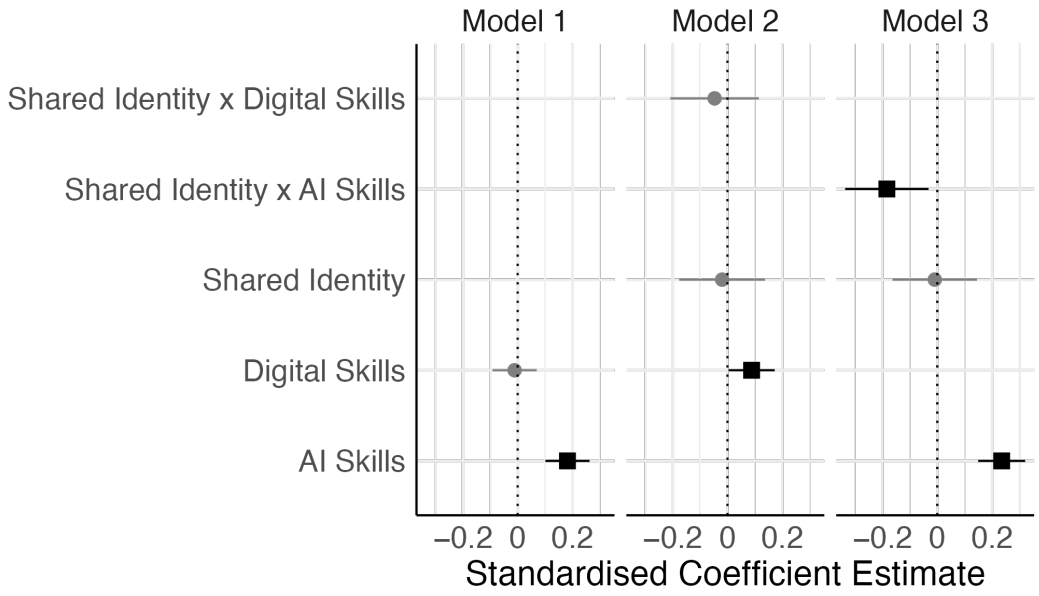

Figure 5. Regressing generative AI perception on shared identity, digital competencies, and their interactions.

Note. Points represent standardized regression coefficients extracted from models 1, 2, and 3, respectively. Lines represent the 95% confidence interval. The dashed line represents the exact null effect. Square points highlighted in bold black are statistically significant at a = .05 and correspond with Table A.5 in the Supplementary Materials.

Discussion

In determining how generative AI could alter political campaign processes, we focused on the ability of campaigns to use these technologies to target out-group voters by altering avatar ethnicity. To operationalize this process, we relied on the creation of a diverse set of synthetic avatars to present the Cape Independence movement's message to potential voters from South Africa's three most populous ethnic groups.

Given the association between the Cape Independence movement and the White population, we hypothesized that the use of non-White avatars to promote the movement would have a disproportionately strong effect in shaping perceptions. Although the results were not statistically significant, the direction and strength of the findings provide *tentative* support for this hypothesis. Specifically, the results suggest that even among individuals outside the Black community, messages delivered by an unexpected source can influence expectations of broader support and potentially shape personal opinions. To further examine the potential for synthetic advertisements to target out-group members, we show that shared identity between the avatars and survey respondents resulted in both increased perceived societal support and personal support for Cape Independence. We find the largest results among Black respondents, which matches expectations and echoes concerns about the potential for generative advertisements to be used to target out-group voters.

The results of these analyses provide further incentive to examine potential solutions capable of mitigating the political influence of generative AI. We find inconclusive evidence about the efficacy of content labeling, which has been proposed as a potential intervention for managing the influence of synthetic media content (Wittenberg et al., 2025). Although the direction of the main analyses suggests that AI labels may slightly reduce the influence of generative advertisements, it is likely not a comprehensive solution.

Alongside these findings, our exploratory analyses indicate that neither participants' ages nor shared ethnic identities influenced their perceptions of AI involvement in the video. However, corroborating existing research in other contexts (Huang et al., 2024; Lu et al., 2024), although the effect sizes were small, digital competency did affect participants' beliefs about whether the video was created using generative AI. Important, AI-specific skills showed a stronger association with AI perception compared to general digital skills, supporting the importance of AI literacy in accurately assessing AI-generated content.

Of more direct interest to the current study, the interaction between shared ethnic identity and AI skills suggests that the relationship between AI skills and the perception of AI-generated content may be influenced by affective involvement. Specifically, for participants who shared an ethnic identity with the avatar, stronger AI skills were associated with a lower likelihood of identifying the content as AI-generated. This implies that familiarity and personal relevance, which are often heightened by shared identity markers, may diminish scrutiny. Put differently, affective involvement, where emotional connections lead to greater engagement, may bias information processing, making individuals more accepting of content that feels personally relevant (Matthes, 2013; Schulz, Wirth, & Müller, 2020). These results can be interpreted through the lens of social identity theory, from which one implication is that identification with media characters can influence interpretation, leading to less critical evaluation of content that resonates with one's identity (Appiah, Knowbloch-Westerwick, & Alter, 2013; Reid, 2012; Trepte & Loy, 2017). This interaction emphasizes the importance of considering both cognitive and affective factors in understanding perceptions of AI-generated material.

These results should be considered in the context of the study's limitations, which include sample size. The overall sample size for this study (n = 875) was primarily determined by resource availability, a common and justifiable constraint in field-based research, particularly in under-resourced contexts (Lakens, 2022; Lenth, 2001). This became relevant as the eligible participants available on Prolific for our study had largely been reached following our set of surveys, which was initially designed to maximize enrollment among Western Cape residents. As such, no a priori power analysis was conducted. Instead, we performed sensitivity analyses to determine the smallest effects that our study was adequately powered to detect. For the analysis presented in Figure 2, which involved a subsample of n = 497 participants, a sensitivity analysis indicated that, assuming a conventional alpha level of 0.05 and 80% power, our design could detect standardized regression coefficients of approximately 0.13 or larger. However, the effects observed in this analysis were smaller (e.g., $\beta \pm 0.06$) and did not reach statistical significance. A similar sensitivity analysis was conducted for the literacy analyses (n = 745), which included two predictors. These were powered to detect effects of approximately $f^2 = 0.06$ or larger, which corresponds to a small- to medium-effect size (Cohen, 1988). Both sets of results suggest that although our design was sensitive to practically meaningful effects, it may not have been capable of reliably detecting very small effects.

Although our findings include several nonsignificant results, we believe these outcomes are still valuable. In contexts where prior meta-analytic estimates are unavailable and foundational data are lacking, particularly for underrepresented populations, studies such as ours provide essential groundwork. The results presented here offer one of the first large-scale empirical examinations of these issues in a South African context and can inform future study designs. Moreover, we believe that transparently reporting null results, interpreted within the bounds of statistical power, contributes meaningfully to a balanced and cumulative scientific record and helps mitigate publication bias (Lakens, 2022).

In addition to these concerns, as the study was conducted in South Africa, the study's location may limit the applicability of the results to states with disparate sociopolitical dynamics. However, this may also represent a strength, as there is a notable scarcity of research in this context, making the findings particularly valuable for understanding how generative AI might influence political processes in underrepresented regions. A second concern relates to our sample, which was not fully representative of the South African population, reducing the generalizability of the results. Finally, the AI literacy scale used in the study focused broadly on knowledge and skills related to interacting with AI-curated content online, rather than specifically targeting generative AI content. This may have affected the accuracy of measuring participants' ability to detect and manage generative outputs.

Future research should build on this work to assess how identity interacts with new forms of generative AI in political contexts. There is a notable gap in the literature about nonoppositional "softfakes" and related efforts to mobilize newer technologies to promote, rather than harm, candidate and campaign reputations. Alongside this work, greater attention should be paid to the development of alternatives to post hoc efforts at content labeling as an intervention to manage the growing influence of synthetic media in political advertisements and campaign materials. Last, our work indicates that conventional digital literacy scales are not well equipped to accurately assess individuals' abilities to identify and manage generative outputs. New scales capable of differentiating between digital and AI literacies are needed to inform the development of effective mitigation strategies.

Conclusion

Deepfakes offer the potential to simulate support from out-group communities while saving time and financial resources and enabling flexibility in the development of responses to ongoing events. As generative technologies become more realistic and widely available, as illustrated here with the case of the secession movement in the Western Cape, it is increasingly necessary for governments to grapple with their influence in political spaces (Chowdhury, 2024).

Our findings demonstrate how advertisements that use generative components to diversify party representatives can effectively target supporters. In addition, by incorporating an AI label within our treatment stimuli, we were able to reaffirm recent studies that highlight the limitations of such labels (Clark & Lewandowsky, 2024). Finally, we found that AI competencies were more effective in predicting accurate assessments of generative content and provide preliminary evidence that shared identity characteristics may lead to lower levels of discernment when assessing generative advertisements.

Policy makers interested in minimizing the political risks posed by generative technologies should be cognizant of how these tools are being used by partisans to distort perceptions of support and inclusivity. Specifically, we hope attention is paid not only to the source of generative ads and election materials but also to the content, which is far more flexible and adaptable to the practices of microtargeting than prior production processes. Moreover, by emphasizing the influence of this comparatively affordable manipulation, the minimal efficacy of labeling, and the contingent effects of AI literacy, we hope this work serves as a reference to develop policies that go beyond mere disclosure to invest in programs that provide citizens with the tools necessary to engage in political arenas that will increasingly involve AI-generated materials.

References

Alon, A. T., Rahimi, I. D., & Tahar, H. (2024). Fighting fake news on social media: A comparative evaluation of digital literacy interventions. Current *Psychology, 43*(19), 17343–17361. doi:10.1007/s12144-024-05668-4

Appiah, O., Knobloch-Westerwick, S., & Alter, S. (2013). Ingroup favoritism and outgroup derogation: Effects of news valence, character race, and recipient race on selective news reading. *Journal of Communication, 63*(3), 517–534. doi:10.1111/jcom.12032

Appiah, O., & Liu, Y.-I. (2009). Reaching the model minority: Ethnic differences in responding to culturally embedded targeted-and non-targeted advertisements. *Journal of Current Issues & Research in Advertising, 31*(1), 27–41. doi:10.1080/10641734.2009.10505255

Barari, S., Munger, K., & Lucas, C. (2021). Political deepfakes are as credible as other fake media and (sometimes) real media. *OSF Preprints*. Retrieved from https://doi.org/10.31219/osf.io/cdfh3

Bayer, J. (2024). Legal implications of using generative AI in the media. *Information & Communications Technology Law, 33*(3), 310–329. doi:10.1080/13600834.2024.2352694

Bennett, B. (2024). Filtering trust: Examining whether disclosing the role of AI changes the impact of viewing AI-generated selfies on body dissatisfaction and trust in technology [Presentation]. Clemson University, Clemson, NC, United States.

Bernard, T. (2024, September). *Poll indicates broad support for regulating ai-generated media related to elections*. Tech Policy Press. Retrieved from https://www.techpolicy.press/pollindicates-broad-support-for-regulating-ai-generated-media-related-to-elections/

Birrer, A., & Just, N. (2024). What we know and don't know about deepfakes: An investigation into the state of the research and regulatory landscape. *New Media & Society,* 1–20. Advance online publication. doi:10.1177/14614448241253138

Botha, J. G., & Pieterse, H. (2020). Fake news and deepfakes: A dangerous threat for 21st century information security. *ICCWS 2020 15th International Conference on Cyber Warfare and Security*, 57–66. doi:10.34190/ICCWS.20.085

Charles, M. (2024). Sign languish: Cape independence party faces uphill battle in garnering signatures for 2024 ballot. Retrieved from https://www.news24.com/news24/southafrica/news/elections-2024-cape-independence-party-struggles-for-ballot-with-only-200-signatures-20240305

Chesney, R., & Citron, D. (2019, January). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs*. Retrieved from https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war

Chia, S. C. (2014). How authoritarian social contexts inform individuals' opinion perception and expression. *International Journal of Public Opinion Research, 26*(3), 384–396. doi:10.1093/ijpor/edt033

Chowdhury, R. (2024). AI-fuelled election campaigns are here—where are the rules? *Nature, 628*(8007), 237–237. doi:10.1038/d41586-024-00995-9

Christopher, N. (2023, July 5). *An Indian politician says scandalous audio clips are AI deepfakes. we had them tested*. Rest of World. Retrieved from https://restofworld.org/2023/indian-politician-leaked-audio-ai-deepfake/

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58*(6), 1015–1036. doi:10.1037/0022-3514.58.6.1015

Clark, S., & Lewandowsky, S. (2024). Seeing is believing: The continued influence of a known AI-generated "deepfake" video. *OSF Preprint*. Retrieved from https://osf.io/t7jfk/download

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2nd ed.). London, UK: Routledge. doi:10.4324/9780203771587

Cooke, D., Edwards, A., Barkoff, S., & Kelly, K. (2024). As good as a coin toss: Human detection of ai-generated images, videos, audio, and audiovisual stimuli. *arXiv Preprint*. Retrieved from https://arxiv.org/abs/2403.16760

Dan, V., Paris, B., Donovan, J., Hameleers, M., Roozenbeek, J., van der Linden, S., & von Sikorski, C. (2021). Visual mis-and disinformation, social media, and democracy. *Journalism & Mass Communication Quarterly, 98*(3), 641–664. doi:10.1177/10776990211035395

Davis, R. (2024, March). *Fact check—is it likely the western cape could become an independent state?* Retrieved from https://www.dailymaverick.co.za/article/2024-03-28-western-cape-independence-fact-check/

de Nadal, L., & Jančárik, P. (2024). Beyond the deepfake hype: AI, democracy, and "the Slovak case." *HKS Misinformation Review, 5*(4), 1–9. doi:10.37016/mr-2020-153

de Rancourt-Raymond, A., & Smaili, N. (2023). The unethical use of deepfakes. *Journal of Financial Crime, 30*(4), 1066–1077. doi:10.1108/JFC-04-2022-0090

de Vries, D., Piotrowski, J. T., & de Vreese, C. H. (2024). *Developing the Digiq: A measure of digital competence*. SSRN Working Paper. Retrieved from https://ssrn.com/abstract=4835479

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics, 26*(1), 69–91. doi:10.1177/1940161220944364

Dogruel, L., Masur, P., & Joeckel, S. (2022). Development and validation of an algorithm literacy scale for internet users. *Communication Methods and Measures, 16*(2), 115–133. doi:10.1080/19312458.2021.1968361

Dunton, C. (2023, May). *Get helpful context with About this image*. Google. Retrieved from https://blog.google/products/search/about-this-image-google-search/

Foos, F. (2024). The use of AI by election campaigns. *OSF*. doi:10.31219/osf.io/zm2r6

Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., . . . & Holz, T. (2024). A representative study on human detection of artificially generated media across countries. *2024 IEEE Symposium on Security and Privacy (SP)*, 55–73. doi:10.1109/SP54263.2024.00159

Gidron, N., Sheffer, L., & Mor, G. (2022). Validating the feeling thermometer as a measure of partisan affect in multi-party systems. *Electoral Studies, 80*, 102542. doi:10.1016/j.electstud.2022.102542

Gregory, S. (2022). Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism. *Journalism, 23*(3), 708–729. doi:10.1177/14648849211060644

Hackenburg, K., & Margetts, H. (2024). Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(24), e2403116121. doi:10.1073/pnas.2403116121

Haleem, A., Javaid, M., Qadri, M. A., Singh, R. P., & Suman, R. (2022). Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks, 3*, 119–132. doi:10.1016/j.ijin.2022.08.005

Helsper, E. J., Scheider, L., van Deursen, A. J., & van Laar, E. (2020). *The youth digital skills indicator: Report on the conceptualisation and development of the ySKILLS digital skills measure*. Katholieke Universiteit Leuven. Retrieved from https://research.utwente.nl/en/publications/the-youth-digital-skills-indicator-report-on-the-conceptualisatio

Henrickson, L. (2023). Artificial intelligence in politics. In S. Coleman & L. Sorensen (Eds.), *Handbook of digital politics* (pp. 242–271). London, UK: Edward Elgar Publishing. doi:10.4337/9781800377585.00026

Hinds, J., & Joinson, A. N. (2018). What demographic attributes do our digital footprints reveal? A systematic review. *PLoS One, 13*(11), e0207112.

Hobbs, R. (2020). *Mind over media: Propaganda education for a digital age*. New York, NY: WW Norton & Company.

Huang, G., Jia, W., & Yu, W. (2024). Media literacy interventions improve resilience to misinformation: A meta-analytic investigation of overall effect and moderating factors. *Communication Research,* 1–28. Advance online publication. doi:10.1177/00936502241288103

Kalla, J. L., & Broockman, D. E. (2020). Reducing exclusionary attitudes through interpersonal conversation: Evidence from three field experiments. *American Political Science Review, 114*(2), 410–425. doi:10.1017/S0003055419000923

Kigwiru, V. K. (2022). *Deepfake technology and elections in Kenya: Can legislation combat the harm posed by deepfakes?* SSRN. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4229272

Knobloch-Westerwick, S., Mothes, C., & Polavin, N. (2020). Confirmation bias, ingroup bias, and negativity bias in selective exposure to political information. *Communication Research, 47*(1), 104–124. doi:10.1177/0093650217719596

Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science, 9*(1), 104–117. doi:10.1017/XPS.2020.37

LaChapelle, C., & Tucker, C. (2023, November 28). *Generative AI in Political Advertising*. Brennan Center for Justice. Retrieved from https://www.brennancenter.org/our-work/research-reports/generative-ai-political-advertising

Langa, J. (2021). Deepfakes, real consequences: Crafting legislation to combat threats posed by deepfakes. *Boston University Law Review, 101*(2), 761–802.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*(3), 187–193.

Long, D., & Magerko, B. (2020). What is AI literacy? competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, *Apr*, 1–16.

Looker, R. (2024, August 19). *Trump falsely implies Taylor Swift endorses him*. BBC News. Retrieved from https://www.bbc.com/news/articles/c5y87l6rx5wo

Lu, C., Hu, B., Bao, M.-M., Wang, C., Bi, C., & Ju, X.-D. (2024). Can media literacy intervention improve fake news credibility assessment? A meta-analysis. *Cyberpsychology, Behavior, and Social Networking, 27*(4), 240–252. doi:10.1089/cyber.2023.0324

Mackin, B. (2024, June 11). *Deepfake ads featuring artificial intelligence-generated images of liberal finance minister Freeland spotted on YouTube*. Retrieved from https://thebreaker.news/business/freeland-deepfakes/

Maphaka, D. (2021). A rhetoric or genuine transformation? an afro-decolonial analysis of democratic alliance economic justice policy. *The Strategic Review for Southern Africa, 43*(2), 40–58. doi:10.35293/srsa.v43i2.767

Martin, Z., Jackson, D., Trauthig, I. K., & Woolley, S. C. (2024, June 6). *Political machines: Understanding the role of AI in the U.S. 2024 elections and beyond*. Center for Media Engagement, University of Texas at Austin. Retrieved from https://mediaengagement.org/research/generative-ai-elections-and-beyond/

Matthes, J. (2013). The affective underpinnings of hostile media perceptions: Exploring the distinct effects of affective and cognitive involvement. *Communication Research, 40*(3), 360–387. doi:10.1177/0093650211420255

Matthews, T., & Kidd, I. J. (2023). The ethics and epistemology of deepfakes. In C. Fox & J. Saunders (Eds.), *The Routledge handbook of philosophy and media ethics* (pp. 342–354). London, UK: Routledge.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, 1–11. doi:10.1016/j.caeai.2021.100041

Papakyriakopoulos, O., Hegelich, S., Shahrezaye, M., & Serrano, J. C. M. (2018). Social media and microtargeting: Political data processing and the consequences for Germany. *Big Data & Society, 5*(2), 1–15. doi:10.1177/205395171881184

Pocol, A., Istead, L., Siu, S., Mokhtari, S., & Kodeiri, S. (2023). Seeing is no longer believing: A survey on the state of deepfakes, AI-generated humans, and other nonveridical media. *Computer Graphics International Conference, 14496*, 427–440. doi:10.1007/978-3-031-50072-5_34

Portelinha, I., & Elcheroth, G. (2016). From marginal to mainstream: The role of perceived social norms in the rise of a far-right movement. *European Journal of Social Psychology, 46*(6), 661–671. doi:10.1002/ejsp.2224

Posel, D. (2001). What's in a name? Racial categorisations under apartheid and their afterlife. *Transformation Durban, 47*, 50–74.

Reid, S. A. (2012). A self-categorization explanation for the hostile media effect. *Journal of Communication, 62*(3), 381–399. doi:10.1111/j.1460-2466.2012.01647.x

Schiff, K. J., Schiff, D. S., & Bueno, N. S. (2025). The liar's dividend: Can politicians claim misinformation to evade accountability? *American Political Science Review, 119*(1), 71–90. doi:10.1017/S0003055423001454

Schulz, A., Wirth, W., & Müller, P. (2020). We are the people and you are fake news: A social identity approach to populist citizens' false consensus and hostile media perceptions. *Communication Research, 47*(2), 201–226. doi:10.1177/0093650218794854

Scott, M. (2024). Moldova fights to free itself from Russia's AI-powered disinformation machine. *POLITICO*. Retrieved from https://www.politico.eu/article/moldova-fights-free-from-russia-ai-power-disinformation-machine-maia-sandu/

Seitz-Wald, A. (2024, February 25). Democratic operative admits to commissioning fake Biden robocall that used AI. *NBC News*. Retrieved from https://www.nbcnews.com/politics/2024-election/democratic-operative-admitscommissioning-fake-biden-robocall-used-ai-rcna140402

Sharma, Y. (2024, February 20). Deepfake democracy: Behind the AI trickery shaping India's 2024 election. *Al Jazeera*. Retrieved from https://www.aljazeera.com/news/2024/2/20/deepfake-democracy-behind-the-ai-trickery-shaping-indias-2024-elections

Shukla, V., & Schneier, B. (2024, June 10). *Indian election was awash in deepfakes—but AI was a net positive for democracy*. Retrieved from https://theconversation.com/indianelection-was-awash-in-deepfakes-but-ai-was-a-net-positive-for-democracy-231795

Simchon, A., Edwards, M., & Lewandowsky, S. (2024). The persuasive effects of political microtargeting in the age of generative artificial intelligence. *PNAS Nexus, 3*(2), 1–5. doi:10.1093/pnasnexus/pgae035

Spivak, R. (2018). "Deepfakes": The newest way to commit one of the oldest crimes. *Georgetown Law Technology Review, 3*, 339–397.

Stuurman, K., & Lachaud, E. (2022). Regulating AI. A label to complete the proposed Act on Artificial Intelligence. SSRN Working Paper. Retrieved from ssrn.com/abstract=3963890

Sule, P. E. (2023). Secession and democratic dictatorships in Africa. In J. O. Chimakonam & I. A. Negedu (Eds.), *African democracy: Impediments, promises, and prospects* (pp. 257–273). London, UK: Bloomsbury Publishing.

Ternovski, J., Kalla, J., & Aronow, P. M. (2022). The negative consequences of informing voters about deepfakes: Evidence from two survey experiments. *Journal of Online Trust and Safety, 1*(2), 1–16. doi:10.54501/jots.v1i2.28

Trepte, S., & Loy, L. S. (2017). Social identity theory and self-categorization theory. In P. Rössler, C. A. Hoffner, & L. van Zoonen (Eds.), *The international encyclopedia of media effects* (pp. 1–13). Malden, MA: John Wiley & Sons. doi:10.1002/9781118783764.wbieme0088

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society, 6*(1), 1–13. doi:10.1177/2056305120903408

Van Deursen, A. J., Helsper, E. J., & Eynon, R. (2016). Development and validation of the internet skills scale (ISS). *Information, Communication & Society, 19*(6), 804–823. doi:10.1080/1369118X.2015.1078834

Wittenberg, C., Epstein, Z., Berinsky, A. J., & Rand, D. G. (2025). Labeling AI-generated media online. *PNAS Nexus, 4*(6), 1–12. doi:10.1093/pnasnexus/pgaf170

Zhang, C., Zhang, C., Zhang, M., & Kweon, I. S. (2023). Text-to-image diffusion models in generative AI: A survey. *arXiv Preprint*. Retrieved from https://arxiv.org/abs/2303.07909