International Review of Social Sciences Research
Volume 5 Issue 3 September 2025
DOI: https://doi.org/10.53378/irssr.353234

Deepfakes, of prey? Precursors and eclipses to the macrocosm Philippine politics

[1]Carie Justine P. Estrellado, [2]Glen B. Millar, [3]Frederick I. Rey & [4]Myla M. Arcinas

Abstract

In the age of disinformation disorder, where a waning in public trust necessitates heightened critical examination of content, deepfakes could further complicate responsive orientations as a contentious mimesis card in political playbook. Grounding on Jean Baudrillard's concept of hyperreality, this paper first briefly examines how these precursors have filled an environment where simulations and simulacra replace authentic experience. Then, through the lens of Habermasian critical theory, it examines how deepfakes "eclipse" the possibility of genuine communicative action. This paper synthesizes that the integrity of public discourse is not an isolated occurrence but rather the culmination of existing gambits and practices from the Philippine political substratum. A priori, one can understand that waiting for a generation to become fully educated on the dangers of deepfakes will take too long, making immediate regulatory intervention essential. Relying solely on media literacy initiatives assumes that Filipinos possess the capacity for sustained critical engagement, yet history shows how political propaganda, sensationalism, and disinformation easily sway public opinion. This vulnerability is exacerbated by the digital ecosystem, where political trolling and echo chambers reinforce manufactured narratives. If traditional propaganda has already distorted perception, deepfakes present an amplified threat, inflating political deception and complicating electoral decision-making. However, given existing political dynamics, the swift implementation and effectiveness of state regulation remains a critical question.

Keywords: *deepfakes, deepfake technology, Artificial Intelligence, Philippine politics, hyperrealism, critical theory*

Article History:

***Received**: April 26, 2025* ***Revised**: July 4, 2025*
***Accepted**: July 7, 2025* ***Published online**: August 1, 2025*

Suggested Citation:
Estrellado, C.J.P., Millar, G.B., Rey, F.I. & Arcinas, M.M. (2025). Deepfakes, of prey? Precursors and eclipses to the macrocosm Philippine politics. *International Review of Social Sciences Research*, 5(3), 26-53. https://doi.org/10.53378/irssr.353234

About the authors:
[1]Corresponding author. Researcher, Tayabas Western Academy. Email: cjpe.twa@gmail.com.
[2]Associate Professor, University of Batangas, Lipa Campus
[3]Associate Professor & Political Sociologist, Researcher Lead & Coordinator, Research Center for Social Sciences and Education, University of Santo Tomas.
[4]Associate Professor, Department of Sociology and Behavioral Sciences, De La Salle University.

1. Introduction

In *Indiana Jones and the Dial of Destiny* (2023), the opening sequence masterfully de-aged Harrison Ford, not through mere visual effects, but through the application of Industrial Light and Magic's FaceSwap, a suite of Artificial Intelligence (AI) referring to images of a younger actor. Also, the movie *Emilia Pérez* (2024) used Respeecher's software to adjust star Karla Sofía Gascón's vocal performance in musical sequences, aiming to make her sound like late 1990s Cher on certain notes. These are a few examples of how AI has penetrated the film industry, redefining the boundaries of visual storytelling and character representation. Now, films brought to life with green screens, elaborate costumes, immersive sound, and breathtaking computer-generated imagery create worlds that feel almost tangible, almost believable. The very AI that fuels this cinematic wonder finds a starkly different application in the political atmosphere—Not sci-fi, of course. But, mesmerizing to see, to feel, to believe.

Imagine AI-generated political satire so convincing; it could sway democratic decisions, beliefs, and truths. Politics, it is a rampage, a credo of locals when elections, and positions of stature politically embedded with will are on the fly. Politics, in the scenes in the Philippines, is not immune to mudslinging candidates, incumbent or not. But what is striking is another element to add to the slingshot: deepfakes.

Globally, deepfakes have infiltrated political spheres with diverse applications and impacts (Gosse & Burkell, 2020; Momeni, 2024; Vaccari & Chadwick, 2020; Wilkerson, 2021). Notable instances include the 2018 face-swapping of leaders like Mauricio Macri and Angela Merkel, and Jordan Peele's collaboration with Buzzfeed on a Barack Obama simulation to raise awareness. The technology led to a Fox affiliate's 2019 dismissal over a Trump fabrication and was used by Extinction Rebellion in 2020 to create a Sophie Wilmès simulation linking deforestation to COVID-19. During the 2020 US presidential campaigns, simulations portrayed Joe Biden's cognitive decline, while the Indian Bharatiya Janata Party employed them for multilingual campaign ads. Bruno Sartori created political parodies in 2020, and 2021 saw the Vovan and Lexus controversy involving alleged impersonations of a Russian opposition figure. In March 2022, a video surfaced online that appeared to show President Zelensky calling on Ukrainian forces to surrender to Russia (Majchrzak, 2023). More recently, a Kamala Harris simulation went viral in 2023, Ron DeSantis's 2023 campaign involved a Trump misrepresentation, and they were widely used for multilingual outreach in India's 2024 state assembly elections.

Also, deepfake photographs are being used to create fictitious online personas, or *sockpuppets*, for disinformation campaigns. One such instance involved a fabricated individual named Oliver Taylor, who published opinion pieces attacking a legal academic, with experts confirming his profile photo as a deepfake. Similarly, in Israel, deepfake images of non-existent people were used on a Facebook page, *Zionist Spring*, to spread right-wing propaganda, with fabricated testimonies broadcast on television despite the inability to verify the individuals' existence.

In the Philippines, a viral video, dubbed the “polvoron” video, alleging President Bongbong Marcos snorted a white powder, was met with skepticism due to its low quality and lack of transparent sourcing. While proponents claimed “authentication,” they provided vague and unverifiable proof. Independent analysis by the Deepfakes Analysis Unit (DAU) revealed facial manipulation, specifically indications of a face swap, in a higher-quality version of the video, further discrediting its authenticity.

2. Literature Review

2.1 Precursors

Early instances of political image manipulation involved placing Abraham Lincoln's face onto the body of another politician, John Calhoun, highlighting the early use of such techniques for political purposes. The phenomenon of “spirit photography” in the late 19th century, where photographers claimed to capture images of ghosts, represents another early form of photographic deception that captured public imagination and sparked debates about the veracity of photographic evidence. While not overtly political, it foreshadowed later concerns.

The 20th century saw further refinement of analog manipulation techniques. Tools like gouache paint, erasers, charcoal sticks, and the airbrush became staples for retouching photographs, leading to the idiom “airbrushed from history”. In the 1920s, Bernarr Macfadden's “composograph” process involved staging news events with actors and then superimposing the faces of the actual individuals involved, illustrating early attempts to create seemingly authentic news visuals through manipulation.

The evolution of video editing also has a rich history preceding the digital era (Dancyger, 2011). Early video editing relied on linear methods, literally involving the physical cutting and splicing of film reels with scissors and tape. The invention of machines like the

Moviola in the 1920s allowed editors to view the film while editing, leading to more precise cuts. Later, electronic editing controllers emerged, offering more sophisticated ways to manipulate footage without physical splicing.

As history recalls, the Soviet Union, under Stalin, heavily employed manual retouching as a tool of political censorship (Hett, 2020) a stark contrast to modern digital face-swapping. Skilled retouchers physically altered photographs, removing individuals deemed enemies of the state, effectively erasing them from official records. This process, primarily "face removal" rather than the replacement achieved with current AI-driven techniques, served to rewrite history and control the political narrative. While both the Soviet practice and modern deepfakes involve manipulating facial imagery, the former relied on manual skill for political purposes, while the latter utilizes complex algorithms for often deceptive or entertainment-focused ends.

In other words, manipulation of visual media is not a new phenomenon. The use of tactics to discredit candidates is a recurring theme. One thing is certain, sophistication connives deception alongside technological advancements (Arugay & Baquisal, 2022; Auburn, 2024; Espina, 2019; Udapa et al., 2021).

2.2 What are Deepfakes?

In techno-centric gaze, the emergence of deepfake technology is a direct consequence of advancements in the field of AI. Deepfakes, described as a portmanteau of “deep learning” and “fake” (Flisak 2024; Lin 2024), refer to synthetic media, which allow for creating hyper-realistic but fake audio or visual content, which stereotypically is associated with depicting people (Vaccari & Chadwick, 2020).

Deepfakes, a term coined in 2017, refer to AI-generated media that convincingly replaces someone's likeness with another's. While the term is relatively new, the technology's roots trace back to the 1990s with CGI advancements and gained momentum in the 2010s due to machine learning and increased computing power. A significant breakthrough occurred in 2014 with the development of Generative Adversarial Networks (GANs), enabling more sophisticated deepfakes. This is a common technique, using two neural networks: one to create fake data and another to check its authenticity. Autoencoders, another method, reconstruct input data to generate new, high-quality deepfakes (Singh et al. 2021).

Cheapfakes, creating fake social media accounts or websites to impersonate politicians or organizations using readily available tools to alter audio recordings, such as changing pitch,

speed, or adding sound effects, cutting out parts of a speech, changing the order of events, or juxtaposing unrelated footage, can mislead information (Hameleers et al., 2025; Le et al., 2024). If cheapfakes purportedly cast a smokescreen over reality, deepfakes, perhaps, are another gamechanger.

2.3 Why do People Believe in Fake News?

According to cognitive psychology and behavioral research, people believe in fake news due to several factors: the *illusory truth effect* (Ahmed et al., 2024; Vasu et al., 2018; Fazio et al., 2019) where repeated exposure increases perceived accuracy; the *source effect* (Pornpitakpan, 2004) where credibility of the source influences belief; the *primary effect* (Vasu et al., 2018) where initial information forms strong opinions; the *ideology effect* (Vasu et al., 2018) where alignment with predispositions drives belief; *confirmation bias* (Vasu et al., 2018) where evidence aligned with beliefs is sought; the *delusion effect* (Bronstein et al., 2019) where delusion-prone individuals accept fake news due to lack of analytical thinking; lack of reflective reasoning where failure to distinguish truth stems from poor reasoning; *bullshit receptivity* (Miller et al., 2023; Pennycook & Rand, 2019) where a belief of nonsense statement to be meaningful; and *overclaiming*, claiming to be "good at spotting fake news." Individuals who overclaim their ability to discern real from fake news are paradoxically more likely to fall for it. Their misplaced confidence makes them less likely to engage in critical thinking needed to identify false information (Pennycook & Rand, 2019; Siar, 2021).

Further, from the detractors' point of tactic, manipulated social media posts and memes are common with emotionally charged content and visual elements (Negrea-Busuioc et al., 2024). Another is *conspiracy theories*, which often tap into public anxieties and distrust of established institutions (Sharma et al., 2023). Even innocuous forms like *satire* and *parody* can be weaponized to intentionally mislead audiences (Hayward, 2024; Zarzosa & Ruvalcaba, 2025). Furthermore, *imposter content*, which involves impersonating trusted sources such as official organizations (Horbyk et al., 2021; Magbanua, 2022; Udapa et al., 2021).

Disinformation campaigns frequently exploit pre-existing societal divisions and biases to amplify their impact. These campaigns often target specific demographic groups or existing political, social, or cultural fault lines to sow discord and deepen societal rifts (Horbyk et al., 2021; Jacobsen & Simpson, 2023). Moreover, the so-called *astroturfing* involves the fabrication of authentic grassroots movements to create a false impression of widespread

public backing for a particular issue (Tufekci, 2017; Mason et al, 2018). This includes staged events like fake protests or the production of misleading advertisements, all designed to attract media attention and online engagement. Generally, emotional manipulation is a common overarching strategy employed in disinformation (Zarzosa & Ruvalcaba, 2025), where tactics like using fear-inducing or anger-provoking content are used to bypass rational thought processes and promote the rapid spread of false narratives.

Constant exposure to disinformation can have significant psychological consequences for individuals. Being bombarded with conflicting and often false information can lead to increased levels of anxiety, confusion, emotional fatigue leading to voting behavior discord (Lutz et al., 2020). This constant state of uncertainty and distrust can also contribute to political polarization by reinforcing existing beliefs (Bronstein et al., 2019; Siar, 2021; Vaccari & Chadwick, 2020) and creating "echo chambers" online, where one can be exposed to information that confirms their own viewpoints, limiting their exposure to diverse perspectives (Jing et al., 2024).

2.4 Of Prey, Digital Facades Targeting the Filipino People

The deception, the deceiver, and the prey. Defacing objection, fake news knows no downfall—Filipinos are the quarry.

The integrity of democratic norms and procedures has been questioned in the run-up to the Philippine elections. Malicious actors have plenty of opportunities to spread lies and steer narratives during campaigns because more and more people are depending on digital platforms for information (Obille et al., 2018).

In this fraught context, the vulnerability of Filipinos to deepfakes stems from a confluence of factors: low digital literacy hinders critical evaluation of online content (Kusumastuti & Nuryani, 2020; Caci et al., 2024), while heavy reliance on social media creates echo chambers and filter bubbles (Espina, 2019). Rapid, unverified information sharing exacerbates the spread of fake news (Espina, 2019). Socio-psychological elements, including political polarization and confirmation bias, emotional engagement, heightened trust in online information (Tuquero, 2022), and a historical context of electoral disinformation (Arugay & Mendoza, 2024), further contribute to this susceptibility. Compounding these issues, the rapid advancement of AI facilitates the creation of increasingly realistic and difficult-to-detect deepfakes.

The accessibility of open-source tools further propelled technology's evolution, with everyday users contributing to its development for both entertainment and malicious purposes. This democratization of technology raised concerns among experts in 2018, leading platforms to implement moderation policies and countries to explore regulatory measures. As of writing, deepfakes have spurred varied responses from nations worldwide. In the US, state laws and a proposed federal act target non-consensual explicit deepfakes. The European Union's AI Act aims to enforce transparency in deepfake creation and dissemination. China's "Deep Synthesis Provisions" emphasize data security and content management. Australia, South Korea, and the UK through their Online Safety Act have enacted laws prohibiting harmful deepfakes, with South Korea notably criminalizing even the any means of distribution of deepfake pornography (MOJ, 2024). Meanwhile, countries like India, Japan, and Singapore are actively exploring regulations and detection technologies to combat the spread of deepfakes. The dissemination of technology and its creations was partly possible due to the simplification of synthesis techniques and the so-called democratization of access to technology (Karlovitz, 2020; Bariach et al., 2024). All this was associated with an increase in disruptive potential, which has accompanied deepfakes since their appearance in 2017.

While the Philippines currently lacks specific legislation directly addressing deepfakes, the Department of Information and Communications Technology (DICT) has indicated that existing laws, such as Article 154 of the Revised Penal Code, can be applied to penalize the erroneous use of deepfakes, particularly in the context of spreading fake news that endangers public order or causes damage to the state's interests. This suggests that while deepfakes are not explicitly addressed in Philippine law, they fall under the broader umbrella of regulations concerning the dissemination of false information.

Deepfake technology is easily accessible to anyone and weaponized to stain reputations, whether for political takedowns or personal vendettas (Dobber et al., 2021; Pawelec, 2024). It has already been exploited to generate fake pornographic material without consent, pushing ethical boundaries into dangerous territory (Łabuz, 2024). But what warrants even graver concern is the near-impossibility of distinguishing the real from the fabricated, even for nations equipped with advanced technological defenses (Rössler et al., 2018; Yu et al., 2021).

The paper highlights two research questions related to:

1. How do the precursors of deepfake technology, when analyzed through the dual sequential lenses of hyperrealism and Habermasian critical theory, illuminate existing imbalances within the Philippine politics?

2. Provide implications of deepfakes as being used (or *could* be used) to manipulate public opinion and reinforce narratives in the context of Philippine politics?

3. Methodology

Unlike other traditional researches conducting experiments and surveys, this research is grounded to critically examine social issues through philosophical lens while advancing arguments, opening a window for further arguments and springing forth data inherent to the philosophical engagement with social issues and interrogating the structures of power and knowledge that reshape social reality. Specifically, this research employs a combined theoretical framework of hyperrealism and critical theory to analyze the impact of deepfakes on the Philippine political arena. Hyperrealism, as theorized by Jean Baudrillard (2001), provides a lens for understanding how deepfakes, as simulations of reality, a junction for a "loss of the real" and the "implosion of meaning" within political adage. Deepfakes, by creating simulacra of political figures and events, can distort public perception and undermine trust in established institutions. This framework allows for an examination of how these simulated realities become more potent than actual events in shaping public opinion (Miller et al., 2023; van Kessel et al., 2025). Furthermore, it enables an exploration of how the proliferation of deepfakes within the hyperreal environment of online media contributes to a crisis of authenticity.

Complementing the hyperrealist perspective, critical theory, particularly Habermas, provides the tools to analyze the power dynamics at play. He even argued that understanding is not a solitary process, but one that is built through social and communicative acts (Geuss, 1981, as cited in Salter, 1983). While Habermas did not explicitly address deepfakes, still it (critical theory) provides a crucial lens for understanding their implication; and examining the advent, dissemination, and reception of deepfakes within the Philippine political context, it only constellates public imaginal narratives. Hence, this combined framework allows for an understanding of truth decay in society and how deepfakes not only distort reality but also

exacerbate existing social inequalities and contribute to the maintenance of dominant ideologies within the political sphere.

4. Discussion

4.1 Deepfakes Through the Lenses of Hyperreality and Critical Theory

Some researchers found deepfakes as tools and not red button for destruction (Broinowski & Martin, 2024; Kerner & Risse, 2021; Liu et al., 2025). The unfounded fear or exaggerating the threat can lead to widespread panic, which can be more harmful than the technology itself. While media effectiveness in prompting action is rooted in its symbolic impact, which can exist independently of factual support (Sharma et al., 2023; van Kessel et al., 2025). The technology itself is neutral, its application depends on the intentions of the user. In this case, the goal is not to inform or educate, but to evoke emotional responses and sway opinions through manufactured representations tending to challenge post humanist thoughts from its risks (Kalpokas & Kalopokiene, 2022). But focusing excessively on the worst-case scenarios this paper holds the ground for deepfakes from its theoretical lacuna.

Jean Baudrillard's concept of hyperreality dates to the nineteenth century when it was considered a provocative and eccentric notion (Biddle & Lea, 2018; Luke, 1991). Hyperrealist artists often utilize photographic images as their primary reference, employing techniques such as grid systems and projectors to transfer these images to their canvases or molds. However, hyperrealism transcends mere photographic replication. Hyperrealism goes the floor as beyond traditional realism. Unlike realism, which reflects life as it (Putnam, 2016), hyperrealism exaggerates elements of reality, making them sharper, more immersive, yet still grounded in real-world cues (King, 2024; Kuryluk, 2023). Authors turn to hyperrealism not to distort facts, but to highlight how human cognition itself navigates heightened simulations. Baudrillard’s concept of the “precession of simulacra” underscores this, simulations no longer just reflect reality; they shape and define it (Luke, 1991). Deepfakes embody this principle, going beyond mere deception to construct entirely fabricated realities. In this sense, hyperrealism does not just confuse reality with illusion, it creates a new layer of perceived truth.

Umberto Eco further explored hyperreality, suggesting that it arises from a desire for reality that paradoxically leads to the fabrication of a false reality, which is then consumed as authentic. The concept of “hyperstition,” developed by the Cybernetic Culture Research Unit, generalizes hyperreality to include "fictional entities that make themselves real," highlighting

the power of ideas to shape reality. In the current digital era, this concept gains prominence as media portrays hyperreal images, making AI powered technologies amplify the world's realism beyond its actuality.

Also, adding another philosophical lense from Jürgen Habermas, a prominent social theorist in the tradition of critical theory, has contributed to understanding of society through his concept of communicative action (Habermas, 1970a, 1989c). This theory posits that society fundamentally operates and evolves through communication that is oriented towards achieving mutual understanding among individuals. At the heart of this concept is communicative rationality, which Habermas defines as communication directed towards reaching, sustaining, and critically reviewing consensus based on the intersubjective recognition of validity claims that can be challenged and defended with reasons (Ashenden & Owen, 1999; Cluley & Parker, 2023). He also opined to dissect scientific modernization of politics wherein it has reduced it to a matter of technical management, neglecting the crucial role of fostering virtuous and actively participating citizens (Habermas, 1973b). However, Fusfield's (1997) critique of Habermas's "argumentative turn" provides an essential lens in understanding the fault line like deepfakes for political discourse. Unlike earlier critical theorists such as Walter Benjamin, Max Horkheimer, and Theodor Adorno, who worked within a declarative rhetorical lineage, Habermas shifts towards Western demonstrative rhetoric, prioritizing oral argumentative discourse over written or indirect forms of communication.

As a recount to counter, Liu et al. (2025) underscore the AI vs. AI arms race in deepfake detection, where advanced neural networks attempt to outmaneuver ever-evolving generative models. Their conjecture, those disruptions in facial movement continuity, particularly in the eyes and mouth, serves as indicators of forgery, demonstrates the technical sophistication required to combat synthetic media (Liu et al., 2025; Qin et al., 2023). But how about the human senses, coiled within hyperrealism? If AI struggles to discern authenticity from forgery, what more the untrained eye? Hyperreality suggests that in a world saturated with simulations, perception is no longer a matter of truth but of believability (Baudrillard, 2001).

If hyperreality has already transformed how people interact with space and media, where do deepfakes enter the picture? If theme parks immerse people in fabricated experiences they willingly embrace**,** it operates as illusions people might not even realize they are consuming (Dobber et al., 2021)**.** While theme parks engineer fantasy for entertainment, deepfakes are injecting synthetic realities into democratic discourse. If hyperreality has taught

us that experience—not truth—shapes belief, then it weaponizes this very phenomenon, crafting political theater where the line between truth and deception is deliberately erased.

Table 1

Simulacra and power: Understanding deepfakes through dual frameworks

Deepfake Technique	Key Components	Nodal implications
Face2Face	Facial manipulation, expression transfer	*Hyperreality*. Obfuscating between genuine expression and fabricated emotion, creating simulated political personas. *Critical Theory*. Reinforces power structures by manipulating public perception and in authentic representation.
Neural Texture Synthesis	Texture transfer, image recoloring	*Hyperreality*. Constructs "realistic" but fabricated scenes, replacing actual evidence with convincing simulations. *Critical Theory*. Undermines truth and justice by creating a reality where evidence is easily manipulated, favoring those with the means to do so.
Lip-sync Deepfake	Audio-visual synchronization, speech synthesis	*Hyperreality*. Creates convincing simulations of speech, potentially dissolving the link between a person's voice and their actual statements. *Critical Theory*. Enables those in power to manipulate narratives, discredit opponents, and control information flow through manufactured "evidence."
Hybrid Deepfake Models	Combination of multiple techniques, adaptive manipulation	*Hyperreality*. Creates complex, multi-layered simulations that are increasingly difficult to distinguish from reality. *Critical Theory*. Amplifies manipulation across various domains, making it harder to discern truth and resist dominant narratives.

Note. Synthesized from Sunkari and Nagesh (2024). *Artificial intelligence for deepfake detection: Systematic review and impact analysis.* The 'Nodal implications' presented in this table offer a generalized overview of the theoretical intersections between deepfake techniques and the lenses of Hyperrealism and Critical Theory. While this framework provides a useful analytical tool, it is essential to recognize that the complex, territorial nature of these dictates, whether intentional or unintentional from the actors, makes distinctions porous between each category. This table should be understood as a starting point for analysis, rather than a definitive classification.

The Filipino public, already immersed in a political rife, may find itself more susceptible to Deepfakes when added to the equation, not because they fail technical scrutiny, but because they align with the narratives people are primed to accept. In this war of perception, detection alone is not enough, when the simulated becomes indistinguishable from reality, the battlefield extends beyond AI into the fragile faculties of human cognition.

The country has witnessed a range of disinformation tactics, each tailored to serve specific political and economic agendas (Cabañes, 2022). Hyperpartisan content, often spread through fake accounts, trolling, and inflammatory speech, thrives in the digital sphere, fueled by informal sector workers, strategists, and political figures who capitalize on public sentiment. Meanwhile, rent-seeking disinformation, orchestrated through advertising and firms, has enabled state actors and corporate entities to shape narratives for their benefit (Cabañes, 2022; Culloty & Suiter, 2021; Lanuza & Arguelles, 2022). Additionally, attention-hacking strategies, such as clickbait-driven content, exploit the Filipino public's heavy reliance on social media, further expanding between fact and fiction (Arguelles & Lanuza, 2020). Unlike other Southeast Asian countries where media is tightly controlled by the state, the Philippines operates within a partly free media environment, where private ownership grants some level of journalistic independence (Lanuza & Arguelles, 2022).

If Habermas's model presupposes a public sphere where rational discourse is paramount, then deepfakes represent an epistemological rupture, challenging the very conditions under which truth claims are assessed. The hyperreal aesthetics of deepfakes blur the boundaries between framing and trolling, intensifying polarization and reinforcing political allegiances, representation and reality, undermining the possibility of rational-critical debate—a central tenet of Habermasian theory. However, this has not shielded it from the influence of partisan actors, nor has it curbed the rampant spread of digital disinformation. With social media serving as both an information source and a battleground for propaganda (Arugay & Baquisal, 2022), deepfakes and other manipulative content pose an even greater threat, making truth a volatile and easily malleable commodity in the country's political discourse.

Deepfakes present a paradoxical technological development, within the entertainment industry while simultaneously posing a severe threat to the integrity of news media (Lundberg & Mozelius, 2024; Wahl-Jorgensen & Carlson, 2021; Waisbord, 2018). Conversely, it challenges the principle of trust. While this threat is acute for vulnerable populations, it also presents an opportunity for news organizations to reinforce their role as arbiters of truth through rigorous source criticism and authenticity verification (Cabañes, 2022; Jacobsen & Simpson, 2023; Magbanua, 2022).

The concept of "eclipses" in Philippine politics, as interpreted in this paper, refer to periods marked by a decline in rational public discourse and democratic ideals. This metaphor

draws on the traditional understanding of eclipses in Philippine mythology, where a mythical creature like the *Bakunawa* was believed to devour the sun or moon, causing darkness and fear.

Table 2

The Habermasian lens on deepfake impact

Concepts/Critiques	Habermasian Lens
Fragility of the Public Sphere & Communicative Action	Deepfakes undermine the conditions for rational discourse, dismantling the potential for genuine public opinion formation. The capacity for citizens to collaboratively construct shared understandings is corroded, leading to a fragmented communication space.
Colonization of the Lifeworld	The intrusion of deepfake technology represents a powerful force colonizing the lifeworld (cf. Habermasian concept of systems). Everyday social interactions are poisoned by uncertainty, where even authentic communication can be questioned as artifice. This disrupts the trust necessary for social cohesion and meaningful political engagement.
Critique of Instrumental Reason	Deepfakes embody the unrestrained application of instrumental reason, where technological means are deployed for purposes of control and manipulation, overriding the values of truth and authentic communication. Where power dictates reality.
Inadequacy of Education Alone	Merely equipping individuals with media literacy overlooks the inherent power asymmetries present in the post-truth environment. Deepfakes illustrate that even ostensibly rational individuals can be manipulated in an information ecosystem weaponized by propaganda and engineered narratives. It shows that personal autonomy is insufficient when facing systemic deception.
Critique of Popper's "Open Society" (Arbitrary Preferences)	Deepfakes expose the vulnerability inherent in a system reliant on individual preferences lacking a foundation of reasoned public deliberation. An environment dominated by manipulated information allows powerful actors to strategically exploit these preferences, undermining open societies.

The overwhelming presence of hyperreal disinformation can eclipse rational thought as overlapping pieces. Several instances in Philippine political history illustrate the interplay between hyperrealism and the bulwark of the public sphere. The 2016 presidential race, for example, was heavily influenced by social media, where pro-Duterte propaganda and disinformation spread rapidly. Emotionally engaging, frequent unverified content contributed significantly to Rodrigo Duterte’s victory (Neilson & Ortiga, 2022; Sinpeng et al., 2020). This content, presented in a relatable manner, exemplifies hyperrealistic elements that resonated with a segment of the population. The subsequent attacks on media outlets critical of the

Duterte administration and the tagging of opposition figures as communist sympathizers demonstrate an assault on the Habermasian ideal of an inclusive and open public sphere. Similarly, the 2022 elections saw widespread disinformation campaigns, including the glorification of Ferdinand Marcos Sr.'s authoritarian past and negative messaging against his political opponents. Narratives portraying a "golden age" under Marcos Sr., often relying on nostalgia and downplaying human rights abuses (Ong, 2022), represent a hyperreal construction of history. The use of "troll farms" and social media influencers to amplify these narratives and attack dissenting voices further highlights the manipulation of the information environment, hindering rational political engagement and eroding trust in factual accounts (Robles, 2024).

4.2 Ammunition in the Philippine Troll Game

What awaits in the battlefield of trolls? A new arsenal. Spin doctors are not just playing the game—they are rewriting the rules. Essentially, trolling associated the online expression of everyday sadism (Buckels et al., 2014). Positioning the nation as "Patient Zero" in this global crisis (Walker, 2022). There is a direct correlation between the growth of troll farms and the spread of disinformation. Trolls in the Philippines, for instance, often operate within organized networks as online trolling often intersects with political discourse, amplifying political narratives and silencing dissenting voices through memes and online harassment to shape narratives (Karpan, 2018). The aesthetic of transgression is often tied to political allegiances (Cabañes & Cornelio, 2017; Neilson & Ortiga, 2022; Walker, 2022).

Deepfakes are strategically employed by trolls for various malicious purposes. Firstly, to magnify disinformation, trolls utilize strategic dissemination on social media, create echo chambers to reinforce biases, and deploy bots and fake accounts to simulate widespread support (Baloglu, 2021). Secondly, for political manipulation, deepfakes are used in character assassination to damage reputations, sow discord among political factions, and facilitate foreign influence operations (Jacobsen & Simpson, 2023; Sharma et al., 2023). Lastly, for financial gain, trolls generate clickbait and ad revenue through deepfakes and engage in paid disinformation campaigns for political or other entities.

Is democracy being at stake? As people retreat into echo chambers and information silos, where their existing biases are reinforced and dissenting voices are silenced, the integrity of electoral processes is profoundly at stake. The problem was not just a few rogue actors, but

a systemic failure. Politics is not to be blamed, but politicians are. In addition to the rapid advancement of AI technology had outpaced regulatory frameworks. The consequences of unchecked deepfake proliferation, from political destabilization to social unrest, could be severe.

4.3 Contesting Normative Assumptions?

When a deepfake can replace a person's face or voice, can an individual truly trust what is seen and heard? Can it be certain that the actions one perceives, even those of real individuals, are not influenced by unseen forces? The onus falls on citizens to become digitally informed and savvy. However, this is not an individual responsibility. It requires a concerted effort from educational institutions, media organizations, and governments to equip citizens with the critical thinking skills and media literacy necessary to navigate the complex information landscape. The challenge is not simply to identify and debunk deepfakes, but to foster a culture of critical engagement with digital media (Jacobsen & Simpson, 2023). This includes promoting media literacy education (Wahl-Jorgensen & Carlson, 2021).

There are no symptoms without diagnosing the ill-aspect of social thinking. That is why education is the forefront to combat deepfakes. To lay a strong foundation, teaching the young people to be analytical from a young age is essential. According to cognitive psychology, a lack of critical and reflective thinking skills is associated with a greater likelihood of falling for fake news (Beauvais, 2022). Thus, it helps to teach kids the fundamentals of digital intelligence and to help them develop critical thinking abilities. Both at home and in school, this ought to start early. However, it is difficult to make up for it and wait for the following generation to completely escape the trap.

4.4 Education [From the Contrary] or a Decisive Swift Regulation?

Advancing the argument, education, by its nature, is a gradual process focused on building awareness and promoting long-term behavioral change. It equips individuals with the tools to discern manipulated content such as critical thinking, but its impact is slow and dispersed due to epistemological understanding (Dwyer, 2023; Leibovitch et al., 2025). In the fast-paced digitization, where deepfakes can spread like wildfire across social media platforms within minutes, this slow responsiveness proves woefully inadequate.

The core challenge in deepfake lies not on the semantic nature of technology but the ability to create highly realistic simulations, indistinguishable from genuine content to the untrained eye, exploits the cognitive vulnerabilities of individuals. While educational programs can raise awareness about these manipulations (Caci et al., 2024; Chemerys, 2024), they cannot guarantee every Filipino critical faculty to identify them in real-time. Moreover, the sheer volume of information circulating online overwhelms even the most vigilant users (Wang, 2022). Education alone cannot stem the tide of misinformation when deepfakes are strategically designed to exploit emotional biases and reinforce existing beliefs. Furthermore, education's impact is often limited by the digital divide. Access to quality education and media literacy resources is not evenly distributed (Roberts & Hernandez, 2019), leaving vulnerable populations susceptible to manipulation (Mason et al., 2018). This creates a fertile ground for deepfakes to thrive, particularly in societies with low levels of digital literacy. In such contexts, the slow and gradual nature of education becomes a significant liability.

In contrast, regulations offer a more immediate and proactive approach, such as content moderation protocols, algorithmic transparency requirements, and legal frameworks that hold creators and distributors of deepfakes accountable, allowing for the disruption of manipulated content before it reaches a critical mass. Regulations can act as a first line of defense, slowing down the dissemination of deepfakes and mitigating their potential for harm. This is not to say that education is irrelevant; rather, it should be seen as a complementary strategy. This is based on the concept of the “ideal speech situation” of Habermas where a normative framework for evaluating existing power structures positing a hypothetical scenario where communication is free from coercion and manipulation.

However, and this is critical, governments cannot simply cry “fire” and then stand idly by. They must demonstrate competence, transparency, and a genuine commitment to truth. Regulations must be carefully crafted, balancing the need for control with the preservation of free speech (Barton & Piston, 2022). If governments themselves are perceived as untrustworthy or incompetent, their efforts to combat disinformation will only fuel the flames of distrust. The urgency of the situation demands a transdisciplinary approach. But combatting deepfakes through education is not easy task, it requires quick responsive, thus the government should be the foreground.

5. Conclusion

Deepfakes did not happen by chance, and as they become hazy, with the distinction between modified and real content confounds, the impact on society is a growing concern. While education can help future generations, still it has limitations in bridging the learning gap. Like a stitch in time, government intervention, could proactively regulate deepfakes. However, such intervention may raise concerns about government competence if authorities prove incapable of effective action.

As fire cannot exist without oxygen, deepfakes rely on the machinery of creation and distribution, even funding. To curb this "information disorder" society must consider measures to restrain its spread. Yet, deepfakes are not merely a technological pursuit but a political arsenal, demanding critical thinking and transparency from those in power and the public alike. It is a struggle to reclaim the integrity of the political state, to restore genuine communication, and to hold disinformants accountable. Only then can one hope to dispel the hyperreal stage and reclaim a politics grounded in truth and public service.

6. Recommendations

Future research avenues may prioritize the implementation of robust, multi-stakeholder frameworks blending technological countermeasures with legal and ethical guidelines, ensuring accountability, with clear mechanisms for recourse, while safeguarding free expression. Concurrently, public literacy initiatives must evolve beyond basic education vis-à-vis advanced critical thinking skills, empowering individuals to discern manipulated content independently. Also, further explorations on the socio-political impact of deepfake proliferation in specific regional contexts and investigate the effectiveness of decentralized verification systems in restoring public trust in digital information and misinformation dynamics.

7. Limitations

Habermas' critical theory and the concept of hyperreality presented insights into the precursors of deepfakes, they possess limitations as an interpretive lens. Both frameworks, rooted in structural analysis and communication theory, tend to underemphasize the dynamism of contextual factors (re)shape the reception and effects of deepfakes. The "lifeworld," as

Habermas conceives it, and the constructed reality of hyperrealism, are treated somewhat empirically, failing to fully account for the rapid and unpredictable shifts in social, political, and technological atmospheres. For instance, the political climate in the Philippines, with its episodic trust in institutions and changing social media trends and communications, can alter how deepfakes are perceived and employed. Furthermore, as of writing, the volatile power structures in the Philippines, from the national up to local politics, driven by provocative idolatry among political parties rather than national patriotism, complicate the application of Habermasian analysis. The hyperreality framework also does struggle to account for the shifts in public allegiance and the manipulation of emotions in such a context. Similarly, while Habermas' critical theory reliance on rational discourse is restrained by a political interest where emotional appeals and personality-driven loyalties frequently override reasoned debate. In addition, we recognize their pre-existing biases and standpoints, since they commit to their non-neutral position in addressing disinformation and information manipulation, asserting a commitment to the national interest. Given the advent of deepfake technologies and its nature, this paper is open to critical scrutiny and further scholarly discourse.

Disclosure statement

No competing interests exist in relation to this work.

Funding

This work was fully funded by the Tayabas Western Academy, Quezon, Philippines. The authors acknowledge the support from its Research and Extension Center, led by Director Mark Vincent P. Aranas, and with the approval of school President Mr. Orlando B. Montecillo.

AI Declaration

The authors declare that Grammarly AI was employed to assist with the formatting of the references in this work.

References

Ahmed, S., Bee, A. W. T., Ng, S. W. T., & Masood, M. (2024). Social media news use amplifies the illusory truth effects of viral deepfakes: A cross-national study of eight countries. *Journal of Broadcasting & Electronic Media*, *68*(5), 778–805. https://doi.org/10.1080/08838151.2024.2410783

Arguelles, C.V. and Lanuza, J.M. (2020). *Linking media systems and disinformation: the case of Southeast Asia*. Manila: Consortium on Democracy and Disinformation.

Arugay, A. A., & Baquisal, J. K. A. (2022). Mobilized and polarized: social media and disinformation narratives in the 2022 Philippine elections. *Pacific Affairs*, *95*(3), 549-573. https://doi.org/10.5509/2022953549

Arugay, A. A., & Mendoza, M. E. H. (2024, July 16). Digital autocratisation and electoral disinformation in the Philippines (ISEAS Perspective 2024/53). *ISEAS – Yusof Ishak Institute*. https://www.iseas.edu.sg/articles-commentaries/iseas-perspective/2024-53-digital-autocratisation-and-electoral-disinformation-in-the-philippines-by-aries-a-arugay-maria-elize-h-mendoza/

Ashenden, S., & Owen, D. (1999). *Foucault contra habermas: Recasting the dialogue between genealogy and critical theory* (1st ed.). SAGE Publications. https://doi.org/10.4135/9781446221822

Auburn, L. (2024, October 22). *Are video deepfakes powerful enough to influence political discourse?* University of Rochester (news center).

Baloglu, U. (2021). Trolls, pressure and agenda: The discursive fight on Twitter [currently X] in Turkey. *Media and Communication (Lisboa)*, *9*(4), 39–51. https://doi.org/10.17645/mac.v9i4.4213

Bariach, B., Hogan, B., & McBride, K. (2024). *Faces of the future: How generative AI is redefining likeness and identity in the age of artificial intelligence*. Oxford Internet Institute.

Barton, R., & Piston, S. (2022). Undeserving rich or untrustworthy government? How elite rhetoric erodes support for soaking the rich. *Politics, Groups & Identities*, *10*(5), 729–753. https://doi.org/10.1080/21565503.2021.1884890

Baudrillard, J. (2001). The precession of simulacra. In D. Durham & D. Kellner (Eds.), *Media and cultural studies: Keyworks*. Blackwell Publishers.

Beauvais C. (2022). Fake news: Why do we believe it? *Joint bone spine*, *89*(4), 105371. https://doi.org/10.1016/j.jbspin.2022.105371

Biddle, J. L., & Lea, T. (2018). Hyperrealism and other indigenous forms of 'faking it with the truth.' *Visual Anthropology Review*, *34*(2), 173–188. https://doi.org/10.1111/var.12148

Broinowski, A., & Martin, F. R. (2024). Beyond the deepfake problem: Benefits, risks and regulation of generative AI screen technologies. *Media International Australia Incorporating Culture & Policy*. https://doi.org/10.1177/1329878X241288034

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition, 8*(1), 108–117. https://doi.org/10.1037/h0101832

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, *67*, 97–102. https://doi.org/10.1016/j.paid.2014.01.016

Cabañes, J and Cornelio, J (2017) The rise of trolls in the Philippines (and what we can do about it). In: Curato, N, (ed.) *A Duterte Reader: Critical Essays On Rodrigo Duterte's Early Presidency*. Ateneo de Manila University Press (pp. 231-250).

Cabañes, J. V. A. (2022). The imaginative dimension of digital disinformation: Fake news, political trolling, and the entwined crises of Covid-19 and inter-Asian racism in a postcolonial city. *International Journal of Cultural Studies*, 25(3-4), 428-444. https://doi.org/10.1177/13678779211068533

Caci, B., Giordano, G., Alesi, M., Gentile, A., Agnello, C., Lo Presti, L., La Cascia, M., Ingoglia, S., Inguglia, C., Volpes, A., & Monzani, D. (2024). The public mental representations of deepfake technology: An in-depth qualitative exploration through Quora text data analysis. *PloS One*, *19*(12), e0313605. https://doi.org/10.1371/journal.pone.0313605

Chemerys, H. (2024). Detection and systematization of signs and markers of modifications in media content for the development of a methodology to enhancing critical thinking in the era of deepfakes. *Fizyko-Matematychna Osvita*, *39*(1), 70–77. https://doi.org/10.31110/fmo2024.v39i1-10

Cluley, R., & Parker, M. (2023). Critical theory in use: Organizing the Frankfurt school. *Human Relations (New York)*, *76*(11), 1689–1713. https://doi.org/10.1177/00187267221111219

Culloty, E., & Suiter, J. (2021). *Disinformation and manipulation in digital media: Information pathologies* (1st ed.). Routledge. https://doi.org/10.4324/9781003054252

Dancyger, K. (2011). *The technique of film and video editing: History, theory, and practice* (5th ed.). Routledge. https://doi.org/10.4324/9780240813967

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. https://doi.org/10.1177/1940161220944364

Dwyer C. P. (2023). An evaluative review of barriers to critical thinking in educational and real-world settings. *Journal of Intelligence*, *11*(6), 105. https://doi.org/10.3390/jintelligence11060105

Espina, J. V. (2019). Online disinformation and its impact on Philippine politics: a case study of the 2019 midterm elections. *Journal of Contemporary Southeast Asian Affairs*, 8(1), 56-73. https://doi.org/10.1177/2334041420907504

Fazio, L. K., Rand, D. G., & Pennycook, G. (2019). Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic Bulletin & Review*, *26*, 1705–1710. https://doi.org/10.3758/s13423-019-01651-4

Flattery, T., Miller, C.B. (2024). Deepfakes and dishonesty. *Philos. Technol.* 37, 120. https://doi.org/10.1007/s13347-024-00812-1

Flisak, D. (2024). O niełatwym wytyczaniu granic rzeczywistości przez Ai Act: Zasada transparentności, deepfake. (Original work published in Polish). *Prawo Nowych Technologii*, 1, 43–46. https://doi.org/10.32027/PNT.24.1.7

Fusfield, W. (1997), Communication without constellation? Habermas's argumentative turn in (and Away from) Critical Theory. *Communication Theory*, 7, 301-320. https://doi.org/10.1111/j.1468-2885.1997.tb00155.x

Geuss, R. (1981). *The idea of a critical theory: Habermas and the Frankfurt School*. Cambridge University Press.

Gosse, C., & Burkell, J. (2020). Politics and porn: how news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication*, *37*(5), 497–511. https://doi.org/10.1080/15295036.2020.1832697

Habermas, J. (1970a). *Toward a rational society: Student protest, science, and politics* (J. J. Shapiro, Trans.). Beacon Press. (Original work published 1968).

Habermas, J. (1973b). *Theory and practice* (J. Viertel, Trans.). Beacon Press. (Original work published 1963)

Habermas, J. (1989c). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society* (T. Burger & F. Lawrence, Trans.). MIT Press. (Original work published 1962).

Hameleers, M., van der Meer, T., & Vliegenthart, R. (2025). How persuasive are political cheapfakes disseminated via social media? The effects of out-of-context visual disinformation on message credibility and issue agreement. *Information, Communication & Society*, *28*(1), 61–78. https://doi.org/10.1080/1369118X.2024.2388079

Hayward, T. (2024). The Problem of disinformation: A critical approach. *Social Epistemology*, *39*(1), 1–23. https://doi.org/10.1080/02691728.2024.2346127

Hett, B. C. (2020). *The Nazi menace: Hitler, Churchill, Roosevelt, Stalin, and the road to war* (First edition.). Henry Holt and Company.

Horbyk, R., I. Löfgren, Y. Prymachenko, & C. Soriano. (2021). Fake News as meta-mimesis: imitative genres and storytelling in the Philippines, Brazil, Russia and Ukraine. *The Journal of the Aesthetics of Kitsch, Camp and Mass Culture* 5 (1): 30–54. https://urn.fi/URN:NBN:fi:aalto-2021112410434

Jacobsen, B. N., & Simpson, J. (2023). The tensions of deepfakes. *Information, Communication & Society*, *27*(6), 1095–1109. https://doi.org/10.1080/1369118X.2023.2234980

Jing, E. L., Goodrick, E., Reay, T., & Huq, J.-L. (2024). Issue fields and echo chambers: Increasing field contestation fueled by moral emotions. *Organization Studies*, *45*(12), 1713–1740. https://doi.org/10.1177/01708406241280004

Kalpokas, I., & Kalpokiene, J. (2022). *Deepfakes: A realistic assessment of potentials, risks, and policy regulation* (1st ed.). Springer International Publishing AG. https://doi.org/10.1007/978-3-030-93802-4

Karlovitz, T. J., & Dirsehan, T. (2020). The Democratization of technology – and its limitation. In *Managing Customer Experiences in an Omnichannel World: Melody of Online and*

Offline Environments in the Customer Journey (pp. 13–25). Emerald Publishing Limited. https://doi.org/10.1108/978-1-80043-388-520201004

Karpan, A. (2018). Case studies in troll armies around the world: The Philippines. In *Troll Factories*. Greenhaven Publishing LLC.

Kerner, C., & Risse, M. (2021). Beyond porn and discreditation: epistemic promises and perils of deepfake technology in digital lifeworlds. *Moral Philosophy and Politics*, *8*(1), 81–108. https://doi.org/10.1515/mopp-2020-0024

King, M. J. (2024). Themeatics: The art of hyperreality. *The Journal of Popular Culture*, 1–19. https://doi.org/10.1111/jpcu.13383.

Kuryluk, E. (2023). Hyperrealism. *Art in Translation*, *15*(4), 419–429. https://doi.org/10.1080/17561310.2023.2308147

Kusumastuti, A., & Nuryani, A. (2020). Digital literacy levels in ASEAN (comparative study on ASEAN countries). In *IISS 2019: Proceedings of the 13th International Interdisciplinary Studies Seminar, IISS 2019, 30-31 October 2019, Malang, Indonesia* (p. 269). European Alliance for Innovation.

Łabuz, M. (2024). Deep Fakes and the artificial intelligence act—an important signal or a missed opportunity? *Policy & Internet*, 1–18. https://doi.org/10.1002/poi3.406

Lanuza, J.M.H. & Arguelles, C.V. (2022). Media system incentives for disinformation. In: H. Wasserman, D. Madrid-Morales (eds) *Disinformation in The Global South*. https://doi.org/10.1002/9781119714491.ch9

Le, A.T., Nguyen, M.D., Dao, M.S., Tran, A.D., & Dang-Nguyen, D.T. (2024). TeGA: A text-guided generative-based approach in cheapfake detection. *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 1294–1299. https://doi.org/10.1145/3652583.3657602

Leibovitch, Y. M., Beencke, A., Ellerton, P. J., McBrien, C., Robinson-Taylor, C.-L., & Brown, D. J. (2025). Teachers' (evolving) beliefs about critical thinking education during professional learning: A multi- case study. *Thinking Skills and Creativity*, *56*, 101725. https://doi.org/10.1016/j.tsc.2024.101725

Lin, L., Xinan, H., Yan, J., Xin, W., Feng, D., & Shu, H. (2024). Preserving fairness generalization in deepfake detection. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.48550/arXiv.2402.17229

Liu, Y., Padmanabhan, B., & Viswanathan, S. (2025). *From deception to perception: the surprising benefits of deepfakes for detecting, measuring, and mitigating bias*. *arXiv:2502.11195*. https://doi.org/10.48550/arxiv.2502.11195

Liu, J., Wang, Lina, Wang, Run, Ke, Jianpeng, Ye, Xi, Wu, Y. (2025). Exposing the forgery clues of deepfakes via exploring the inconsistent expression cues, *International Journal of Intelligent Systems*. 7945646. https://doi.org/10.1155/int/7945646

Luke, T. W. (1991). Power and politics in hyperreality: The critical project of Jean Baudrillard. *The Social Science Journal*, *28*(3), 347–367. https://doi.org/10.1016/0362-3319(91)90018-Y

Lundberg, E., & Mozelius, P. (2024). The potential effects of deepfakes on news media and entertainment. *AI & Society*, 40, 2159–2170. https://doi.org/10.1007/s00146-024-02072-1

Lutz, B., Adam, M. T. P., Feuerriegel, S., Pröllochs, N., Neumann, D., Riedl, R., vom Brocke, J., Léger, P.-M., Randolph, A. B., Davis, F. D., & Fischer, T. (2020). Identifying linguistic cues of fake news associated with cognitive and affective processing: Evidence from NeuroIS. In *Information Systems and Neuroscience* (Vol. 43, pp. 16–23). Springer International Publishing AG. https://doi.org/10.1007/978-3-030-60073-0_2

Magbanua, K. S. (2022). An analysis of the legal and ethical implications of online disinformation in the Philippines. *Journal of Public Representative and Society Provision*, *2*(2), 52–55. https://doi.org/10.55885/jprsp.v2i2.201

Majchrzak, A. (2023). Rosyjska dezinformacja i wykorzystanie obrazów generowanych przez sztuczną inteligencję (deepfake) w pierwszym roku inwazji na Ukrainę. *Media - Biznes - Kultura. Dziennikarstwo i komunikacja społeczna*, *1*(14), 73–86. https://doi.org/10.4467/25442554.MBK.23.005.18028

Mason, L. E., Krutka, D. G., & Stoddard, J. D. (2018). Media literacy, democracy, and the challenge of fake news. *The Journal of Media Literacy Education*, 10(2), 1–10. https://doi.org/10.23860/jmle-2018-10-2-1

Miller, E. J., Steward, B. A., Witkower, Z., Sutherland, C. A. M., Krumhuber, E. G., & Dawel, A. (2023). AI hyperrealism: Why AI faces are perceived as more real than human ones. *Psychological Science*, *34*(12), 1390–1403. https://doi.org/10.1177/09567976231207095

Ministry of Justice (MOJ) International Legal Policy Division [Editorial Department]. (2024). Regulating deepfakes in politics. *Recent Trends of Law & Regulation in Korea*, *43*, 9–11.

Momeni, M. (2024). Artificial intelligence and political deepfakes: Shaping citizen perceptions through misinformation. *Journal of Creative Communications*, 20(1), 41-56. https://doi.org/10.1177/09732586241277335

Negrea-Busuioc, E., Ștefăniță, O., & Buf, D.-M. (2024). Romania's first female prime minister's meme-ification: Humor and the trivialization of politics in satirical memes. *Journal of Language and Politics*. https://doi.org/10.1075/jlp.22010.neg

Neilson, T., & Ortiga, K. (2022). Mobs, crowds, and trolls: Theorizing the harassment of journalists in the Philippines. *Digital Journalism*, *11*(10), 1924–1939. https://doi.org/10.1080/21670811.2022.2126990

Obille, K. L. B., Gillet, V., McLeod, J., Willett, P., & Chowdhury, G. (2018). Information behavior and Filipino values: An exploratory study. In *Transforming Digital Worlds* (Vol. 10766, pp. 521–526). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-78105-1_57

Ong, J. C. (2022). Philippine Elections 2022: The dictator's son and the discourse around disinformation. *Contemporary Southeast Asia*, *44*(3), 396–403.

Pawelec, M. (2024). Decent deepfakes? Professional deepfake developers' ethical considerations and their governance potential. *AI Ethics*. https://doi.org/10.1007/s43681-024-00542-2

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, *88*(2), 185–200. https://doi.org/10.1111/jopy.12476

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology, 34*(2), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x.

Putnam, H. (2016). Realism. *Philosophy & Social Criticism*, *42*(2), 117–131. https://doi.org/10.1177/0191453715619959

Qin, W., Zou, B., Li, X., Wang, W., & Ma, H. (2023). Micro-expression spotting with face alignment and optical flow. In *Proceedings of the 31st ACM International Conference on Multimedia*, 9501–9505. https://doi.org/10.1145/3581783.3612853

Roberts, T., & Hernandez, K. (2019). Digital access is not binary: The 5'A's of technology access in the Philippines. *The Electronic Journal of Information Systems in Developing Countries*, *85*(4), e12084. https://doi.org/10.1002/isd2.12084

Robles, A. (2024). Trolls on front line of presidential fallout; Supporters of Marcos and predecessor Duterte turn social media blue in ugly clashes that reflect use of indispensable weapon for rival politicians. *South China Morning Post*.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2018). FaceForensics: A large-scale video dataset for forgery detection in human faces. *arXiv*. https://doi.org/10.48550/arxiv.1803.09179

Salter, S. (1983). The idea of a critical theory: Habermas and the Frankfurt school by Raymond Geuss. Cambridge University Press. *New Blackfriars*, *64*(762), 533–534. https://doi.org/10.1017/S0028428900032030

Sharma, I., Jain, K., Behl, A., Baabdullah, A., Giannakis, M., & Dwivedi, Y. (2023). Examining the motivations of sharing political deepfake videos: the role of political brand hate and moral consciousness. *Internet Research: Electronic Networking Applications and Policy*, *33*(5), 1727–1749. https://doi.org/10.1108/INTR-07-2022-0563

Siar, S. V. (2021). *Fake news, its dangers, and how we can fight it* (Policy Notes No. 2021-06). Philippine Institute for Development Studies.

Singh, R., Sarda, P. & Aggarwal, S. (2021). Demystifying deepfakes using deep learning. In: *Proceedings - 5th International Conference on Computing Methodologies and Communication*, ICCMC 2021, pp 1290–1298, https://doi.org/10.1109/ICCMC51019.2021.9418477

Sinpeng, A., Gueorguiev, D., & Arugay, A. A. (2020). Strong fans, weak campaigns: Social media and Duterte in the 2016 Philippine election. *Journal of East Asian Studies*, *20*(3), 353–374. https://doi.org/10.1017/jea.2020.11

Smith, J. (1992). Reviews: Jürgen Habermas, the structural transformation of the public sphere: An inquiry into a category of bourgeois society (Polity Press/MIT, 1989). *Thesis Eleven*, *31*(1), 182–187. https://doi.org/10.1177/072551369203100115

Sunkari, V., & Nagesh, A. S. (2024). Artificial intelligence for deepfake detection: Systematic review and impact analysis. *IAES International Journal of Artificial Intelligence*, *13*(4), 3786-3792. https://doi.org/10.11591/ijai.v13.i4.pp3786-3792

Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press. https://doi.org/10.25969/mediarep/14848

Tuquero, L. (2022, February 26). *51% of Filipinos find it difficult to spot fake news on media – SWS*. RAPPLER. https://www.rappler.com/nation/sws-survey-fake-news-december-2021/

Udupa, S., Gagliardone, I., & Hervik, P. (2021). *Digital hate: the global conjuncture of extreme speech*. Indiana University Press.

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). https://doi.org/10.1177/20563051209034

van Kessel, C., Manriquez, J. D., & Kline, K. (2025). Baudrillard, hyperreality, and the 'problematic' of (mis/dis)information in social media. *Theory & Research in Social Education*, 1–23. https://doi.org/10.1080/00933104.2024.2439302

Vasu, N., Ang, B., Teo, T-A., Jayakumar, S., Faizal, M., & Ahuya, J. (2018). Human fallibility and cognitive predispositions. In *Fake news: National security in the post-truth era*. S. Rajaratnam School of International Studies, Nanyang Technological University.

Wahl-Jorgensen, K., & Carlson, M. (2021). Conjecturing fearful futures: Journalistic discourses on Deepfakes. *Journalism Practice*, *15*(6), 803–820. https://doi.org/10.1080/17512786.2021.1908838

Waisbord, S. (2018). Truth is what happens to news: On journalism, fake news, and post-truth. *Journalism Studies*, *19*(13), 1866–1878. https://doi.org/10.1080/1461670X.2018.1492881

Walker, T. (2022). Trolls, disinformation make Philippine election coverage a challenge. In *Voice of America News / FIND*. Federal Information & News Dispatch, LLC. https://www.voanews.com/a/trolls-disinformation-make-philippine-election-coverage-a-challeng e/6519577.html

Wang, A. H.-E. (2022). PM me the truth? The conditional effectiveness of fact-checks across social media sites. *social media + society*, *8*(2). https://doi.org/10.1177/20563051221098347

Wilkerson L. (2021). Still waters run deep: The rising concerns of 'deepfake' technology and its influence on democracy and the first amendment. *Missouri Law Review*, 86(1), 407–432.

Yu P., Xia Z., Fei J., & Lu Y. (2021). A survey on deepfake video detection. *IET Biometrics*, 10(6), 607–624. https://doi.org/10.1049/bme2.12031

Zarzosa, J., & Ruvalcaba, C. (2025). Fighting fake news: Building cognitive resistance and mental antibodies through a digital media literacy inoculation intervention. *Journal of Marketing Education*. https://doi.org/10.1177/02734753251319559