Ahmad Jamal

A Jamal, Independent research scholar.

E-mail:
ahmad1304@gmail.com

Anthony Collins

Prof. A Collins, International Centre of Nonviolence, Durban University of Technology
Durban, South Africa; Department of Social Inquiry, La Trobe University
Melbourne, Australia.

E-mail:
a.collins2@latrobe.edu.au

First submission: 1 May 2025

Acceptance: 14 November 2025

Published: 12 December 2025

https://doi.org/10.38140/aa.v57i2.10222

ISSN: 0587-2405

e-ISSN: 2415-0479

Acta Academica • 2025 57(2): 166-189

Propaganda, the public spheres, and the new economic conditions of the digital age

Abstract

The arrival of mobile internet services came with significant promise of a potentially democratic public sphere mediated through digital media technology. However, rather than achieving this promise, these technologies ushered in a new economic age: surveillance capitalism, which revolutionised propaganda and misinformation in ways which pose clear threats to civil order and democratic processes. This paper updates and integrates Jürgen Habermas's model of the public sphere and communicative rationality with the realities of surveillance capitalism's new economic logics and the ways in which they undermine the public sphere. (1) The selective distribution of content by sorting algorithms fosters cultures of insularity and antagonism which (2) fragment and disintegrate the sphere, contributing to the breakdown of public discourse into echo chambers and filter bubbles. Furthermore, (3) sorting algorithms facilitate the evasion of scrutiny as these individually selected messages bypass public attention and criticism. This lack of scrutiny encourages the use of (4) engagement-farming tactics like the use of inflammatory misinformation. This has often led to (5) patterns of migration; wherein users migrate away from the platform to find more corresponding information about the

conspiracy or mythology in question – further breaking down public discourse. Lastly, answering the new revenue model's demand for content often results in (6) content from botnets, troll farms, and misinformation networks being platformed and circulated with little to no repercussions.

Keywords: surveillance capitalism, misinformation, social media, echo chambers, radicalisation

A new economic era

This paper is motivated by the Bell Pottinger/Guptabot disinformation campaign in South Africa, the Cambridge Analytica scandal in the US, and the genocide of the Rohingya in Myanmar that was facilitated through a disinformation campaign on Facebook. The aim is to critically analyse the new technologies and methodologies being harnessed by public relations firms and digital age propagandists. It became clear that analysis of these technologies and methodologies was about much more than modern propaganda, and required the conceptualisation of the evolution of late-stage capitalism into a new form: Surveillance Capitalism, a new economic age which required its own set of cultural critique.

The term refers to the saturation of human society by digital devices and interfaces which surveil users in order to map their subjectivity and behaviour (Zuboff 2019). These behavioural insights and understandings of the user are processed into predictions of their behaviour which are then sold to advertisers, PR firms, and propagandists – enabling the critical contextual conditions from which these contemporary methodologies of social influence arose; where a user's subjective information is used to customise messaging and propaganda that targets them individually (Zuboff 2019).

These propagandists are not the only subject of this paper. We argue for a reconceptualisation of the internet and the digital age to better understand its benefits and dangers. This includes a focus on the role of the internet and social media as a potentially democratic public sphere – and the ways in which Surveillance Capitalism and the new economic logics of the digital age manipulate these spheres into zones of economic interest which prioritise their own objectives. This paper argues that these elements are *harmful* to the public sphere and contribute to the erosion of critical democratic processes and the loss of trust in critical democratic institutions. At the same time,

we have witnessed the rise in *anti-publics* online, from echo chambers and forums that have inspired mass shooters and white supremacist terrorism, to the leaders of the 'intellectual dark web' who promote anti-democratic movements, conspiracy theories, and ideologies. These are groups whose outward goals reflect antidemocratic positions and produce engagement through misinformation, stereotypes, fallacies and biases. This anti-disourse goes beyond contrarian or counter discourse; for example, trolling breaks down attempts at discourse, while bots are simulations of individuals simulating discourse. Lastly, with the second election of Donald Trump, the rise of Elon Musk and X to political leadership roles, and the assemblage of the tech industry tycoons under the Trump banner, the time is late for us to reconsider and reevaluate our relationship with the digital world.

The Public Sphere

Jurgen Habermas's concept of the Public Sphere envisions an idealistic arena for public discourse where private citizens could meet and discuss and voice their opinions on matters of public concern (Habermas 1989). From salons to coffee houses, these were spaces where private citizens could discuss matters of public health, law and governance, and ethics and morality (Habermas 1989). Habermas saw it as a space where individuals could exercise communicative rationality to voice their concerns and settle debates on matters that affected their lives (Habermas 1989). For Habermas, rationality was something that could be achieved between people through successful communication. In a discussion between various people with various viewpoints, observing the necessary conditions for healthy discourse and dialogue can allow these people to successfully debate and argue, to collectively determine a resolution to their debate, and arrive at a new rationality and understanding together. Habermas's ideal conception of the public sphere saw it as an arena where such exchanges of communicative rationality could occur.

Communicative rationality thus becomes a requisite for a healthy society; it allows individuals and communities to discuss, debate, and collectively determine their needs and resolutions to their concerns. This dynamic allows them to champion their concerns and perspectives in relation to the concerns of the upper-class elites, creating a foil against the excesses of power and against totalitarian and fascist regimes.

Habermas uses the term *Bourgeois Public Sphere* – indicating the particular class of citizenry (bourgeois) that was able to access the sphere: the emerging class of merchants, professionals and intellectuals who gained cultural influence in the 18th century. In seeking autonomy from state structures, these bourgeois intellectuals and tradesmen sought spaces for public discussion and critique of local politics and policies, and other public matters that had an impact on the lives of individuals. However, this also meant that certain members of the public were not able to access the sphere owing to class, gender, literacy levels, and more. These elements limited access to the sphere, resulting in the public sphere being one that was not truly democratic and public.

Furthermore, he saw the public spheres themselves as being manipulated by powerful interests like the corporations that turned newspaper media into advertising media (Habermas 1989). These strategic interests undermined the ability of the sphere to facilitate communicative rationality. Instead, these interests support a different rationality – *strategic rationality*, wherein the strategic interests of private elites and corporations are exercised to manipulate the discourse within the sphere – from advertising to public relations.

As many scholars and critics point out, it becomes necessary to question the relevance of Habermas's idealistic model in the world of today (Amiradakis 2019, Mylonas 2023). Habermas explicitly contextualises it within a particular historical period with its own social and cultural dynamics (Habermas 1989, Amiradakis 2019).

The work of recent scholars highlights two key points of Habermas's thesis which mark its relevance for contemporary scholarship: (i) a particular set of social, historical and economic dynamics gave rise to the bourgeois public sphere – even if it was short lived; (ii) this sphere began to erode under the weight of strategic interests and manipulations, from advertisers and marketers to propagandists (Amiradakis 2019). Both these points give us a means of conceptualising the growth and decline of the sphere.

To this, the study adds a third point regarding the relevance of Habermas's model today: (iii) the internet. The arrival of the internet was regarded as the fulfilment of a Habermasian-Weberian ideal type of public sphere – one that went beyond the limitations of the bourgeois public sphere to one that was more democratic and accessible to all (Dahlgren 2005, Zuboff 2019).

Habermas's exploration of the growth and decline of the public sphere gives a means to conceptualise the internet – particularly under surveillance capitalism – creating a parallel to Habermas's own argument in which particular conditions permitted the public sphere to emerge – yet it soon declined under the weight of strategic interests. Similarly, specific technological conditions allowed for the digital public sphere to emerge, yet the discovery of surveillance revenues and surplus data fomented a set of strategic ambitions that undermine the digital public sphere and led to its decline.

Google's discovery of the revenue potential of the behavioural surplus data it had been gathering through its service led to the dawn of surveillance capitalism (Zuboff 2019). Initially considered a waste by-product of these digital services, in Google's search for a profitable model, its lead engineers realised that this data could be processed into profitable market insights based on analysis of the online engagements of each individual. The more data accrued, the greater the ability to predict the behaviour of each online individual. Social networks and websites were quick to follow, opening up monetisation opportunities beyond selling advertising space. Now they were able to sell user data, and insights into their users' engagements, interests, preferences, and online behaviour (Zuboff 2019).

This heralded a new gold rush, leading to the entry of other industries into these new markets; all seeking to digitise elements of their products and services in order to surveil their customers and accrue behavioural information – from smart televisions to Barbie Dolls fitted with recording devices and digital interfaces. This created the underlying economic logics that motivate the machinery of the digital age. Consequently, rather than these networks being arenas for democratic discourse mediated by technology, these are instead zones of economic interest which mediate a communications infrastructure that prioritises its own economic objectives.

Fake news, clickbait articles, and polarising content aimed at engaging an audience rather than cultivating discourse have all arisen as new modes of strategic rationality; while technologically advanced public relations firms and cyberwarfare firms have developed entirely new arsenals of strategic tools. Bots, for example, represent a new kind of strategic threat intended to resemble real private citizens engaging in organic discourse within the sphere but are instead facades of human beings veiling and asserting strategic interests. Existing as simulations of individuals – strategic simulacra – bots have

been used in everything from electioneering campaigns, to inciting violence and coups, and even to direct a genocide (Gu, Kropotov and Yarochkin 2017, Strydom 2017, Mozur 2018, Wylie 2019, CABC 2021, Andrzejewski 2023).

Thus, while the internet provided a greater degree of access to the public sphere for working-class proletariat voices to be expressed, it also allowed for a greater number of individuals to be surveilled and monitored and their data extracted and processed into predictions of their behaviour, thus enabling these strategic interest groups to determine how to influence that particular individual. This essentially allows for the subjective experiences of end-users to be mapped and exploited by elite interests (Wylie 2019, Zuboff 2019).

How Surveillance Capitalism harms the Public Sphere

The content crisis

Content is the fuel that drives the machinery of Surveillance Capitalism. Content motivates engagement with the platform, and allows the revenue processes of Surveillance Capitalism to be operationalised – from advertising exposure and conversions, to surveillance, extraction, and prediction. Like the culture industry envisioned by Theodore Adorno and Max Horkheimer – an all-encompassing industry producing popular cultural products (films, series, games, fashion) that pacify the masses into passive consumerism – this dependence on content has created a new culture industry of individual content creators, influencers, gurus, thought-leaders and more. This has led to the evolution of social networks into social media – where they have become entertainment-on-demand platforms and essential resources for keeping individuals socially connected. But, for reasons we will explain, this dependence has also resulted in these public spheres becoming saturated with inflammatory misinformation and conspiracy theories.

Section 23 of the Communications Decency Act of 1996 is the central piece of US legislation that governs content hosted by these platforms. It does not hold the platforms liable for the content they host and regards them as intermediaries as opposed to publishers. This grants the platforms a degree of immunity (Wylie 2019, Zuboff 2019). Actions taken to remove problematic content typically only occurs when content threatens the flows of engagement, or when it goes against the ideological standing of the platform (Zuboff 2019). This can be seen in the ways content featuring Palestine is circulated much less

on Western social media platforms than similar content featuring Ukraine. This ideological dynamic was a key motivator in the reinstatement of previously banned accounts on Twitter when it was rebranded as X under Elon Musk's takeover (Ahmed 2023). Content seen as harmful from a more liberal social perspective was no longer viewed as problematic from Musk's perspective, but rather as lucrative engagement. Andrew Tate's account alone accounted for $13 million in annual advertising revenue (Ahmed 2023).

In more recent times, Mark Zuckerberg has announced the removal of fact-checkers from Meta, which has since seen a marked rise in misinformation across the platform. Zuckerberg, along with Amazon's Jeff Bezos, has joined the ranks of other tech-industry titans like Musk and Pieter Thiel in aligning with the Trump administration's rhetoric and attitude towards social media. This political front positions itself as opposed to the 'woke' suppression of hate speech against vulnerable groups and socially harmful disinformation, such as false medical claims. Musk himself came to run the Department of Government Efficiency (DOGE); and has repeatedly spread misinformation and conspiracy theories – now, as a lead figure in global politics.

The breakdown of public discourse

Sorting algorithms determine which user gets which piece of content – this means that each user sees different content, creating environments of informational asymmetry (Gorton 2016). This dramatically undermines the ability of the digital public spheres to host collective public discourse, as individuals are fragmented into disparate engagement networks, which is harmful to social cohesion and democratic processes (Gorton 2016). This is true of all social networks that use recommendation engines and sorting algorithms, as well as news media aggregators that filter and select news content for their audiences.

Additionally, through 'ad managers' on platforms like Facebook, Instagram, and TikTok which offer marketers services like 'custom audiences', a marketer or PR agent can handpick the type of individual best suited for their messaging – down to specific subjective and psychographic variables. This is known as *psychographic microtargeting*. As opposed to traditionally used demographic variables, psychographics focuses on the subjectivity of particular individuals which is profiled through surveillance of their digital engagements (through both online platforms as well as devices with digital interfaces and components).

Microtargeting contributes to harmful climates of insularity – as content is only shared with particular individuals and groups within the public sphere; often forming insular echo chambers and filter bubbles. The 'filter bubble' effect was explained by internet activist and early warner of the perils of the internet Eli Pariser in his 2011 book *The Filter Bubble: What the Internet is Hiding From You*. It refers to a stituation where recommendation engines create a state of 'intellectual' isolation by only selecting content that will resonate with the specific user – and isolating them from content and perspectives that are different to theirs (Pariser 2011).

Furthermore, these services provide analytical features that allow the marketer to see how their messaging is performing with the various groups they have defined. This can result in the marketer recognising and reinforcing patterns of engagement that are successful. This type of deliberate psychographic reinforcement can contribute to the development of biases and stereotypes, as well as providing confirmation for existing biases and stereotypes. For example, racist content routinely either presents stereotypes of particular populations, or provides information that supports and confirms biases and stereotypes of that population. This type of inflammatory content has proven popular among content creators who recognise that antagonistic content is conducive to engagement. Within the context of filter bubbles, this results in the formation of filter bubbles and echo chambers that are saturated with misinformation and fragment the public spheres along these emotionally antagonistic divides. This is sometimes referred to as *affective polarisation* (Gorton 2016). Platforms can themselves determine the degree of circulation that content receives – suppressing the circulation of content related to politically contested issues like Palestine. A recent example is how Motaz Azaiza's journalism content was maliciously falsely flagged as pornographic and thus received much lesser circulation (Reuters 2024). *This reflects deliberate ideological control over the public sphere*. This is Habermas's strategic rationality not simply applied within the sphere to manipulate communication – but applied *by* the sphere itself.

Evasion of scrutiny

Crucially, microtargeted messages are able to avoid scrutiny; whether from other members of the public, or from academics, scholars, politicians, and journalists. This encourages advertisers and campaigners to introduce embellishments into their messages; to use inflammatory or even false information in their communications.

Consequently, even content that is distributed on a mainstream network like Facebook has a remarkably high potential to evade scrutiny as content becomes individually selected for particular people. This is a critical break in its potential as a healthy public sphere. During the South African State Capture saga, then-President Jacob Zuma was on trial for corruption with particular focus on his relationship with the wealthy Gupta family from India. The term 'State Capture' had been developed in order to represent the nature of the political scandal – wherein the state had fallen under the control of an oligarchy of foreign private citizens. Faced with the loss of support from his key demographics, Zuma and the Guptas engaged with several figures and firms in order to spin doctor these allegations of corruption and State Capture. This led to them contracting with public relations and electioneering firm Bell Pottinger, as well as contracting with service providers from India who developed a disinformation network of accounts across Twitter and Facebook.

This disinformation network – which came to be known as the 'RET Network' (Radical Economic Transformation, after the since-removed faction of the same name within Zuma's ANC government) – conducted an 'astroturfing' campaign of deliberately constructed inflammatory and polarising political claims claiming to represent a popular movement in South Africa's online public spheres (CABC 2021). These communications received critical scrutiny only once it was too late, after the messaging and its ideologies had already been distributed to the initial target audience – from where it received further amplification through shares and retweets (CABC 2021). Additionally, fake accounts impersonating real people are also able to evade scrutiny. These accounts were used to engage with and amplify the narratives being peddled by these propagandists, creating the 'illusion of organic support' for these narratives by what appeared to be ordinary South African citizens (CABC 2021).

This evasive communicative landscape can also allow bad actors to *advertise* on these platforms while avoiding scrutiny. The 'RET Network' placed adverts for its attack-site, 'WMC Leaks', on Facebook, *using Facebook's own advertising services* (below, Strydom 2017). The ad is for an article accusing Johann Rupert of corruption. This was the strategy of the network; to place the blame for the country's woes on white elites within South Africa through labels and conspiracy theories like Stellenbosch Mafia and White Monopoly Capital (WMC – which is where the website gets its name) – while spin-doctoring the focus away from Jacob Zuma and the Guptas.

Figure 1: WMC Leaks advert on Facebook, demonstrating how the social networking giant was culpable (even if unintentional); earning ad revenue by placing ads for the misinformation campaign.

Source: (Strydom 2017)

Additionally, the teams of moderators that work for these platforms are hindered by the variety of languages and dialects spoken within some regions. This allows inflammatory and destructive content to evade scrutiny from those hired to scrutinise for hateful content. In Myanmar, Facebook's moderation team could not keep up with the variety of dialects spoken; allowing for hateful content to proliferate – ultimately resulting in a genocide and the largest mass migration in recent history (Mozur 2018).

Native Advertising (also known as *Programmatic Advertising*) services like MGID allow for the placement of ads in the online spaces and websites that the specific target consumer will visit. However, this creates a situation where the brand or product being advertised has no knowledge of where their adverts will end up (Le Roux 2018). This can result in major brands offering unscrutinised support to fake news networks by monetising their content with advertising revenue (Le Roux 2018). A local investigation into a burgeoning network of associated fake news sites here in South Africa demonstrated how native advertising services resulted in major retailers like Takealot advertising on these fake news websites. Ironically, the investigation was conducted by News24, the sister company of Takealot. Crucially, this type of microtargeted advertising service places adverts without the knowledge of the brand being advertised (Le Roux 2018).

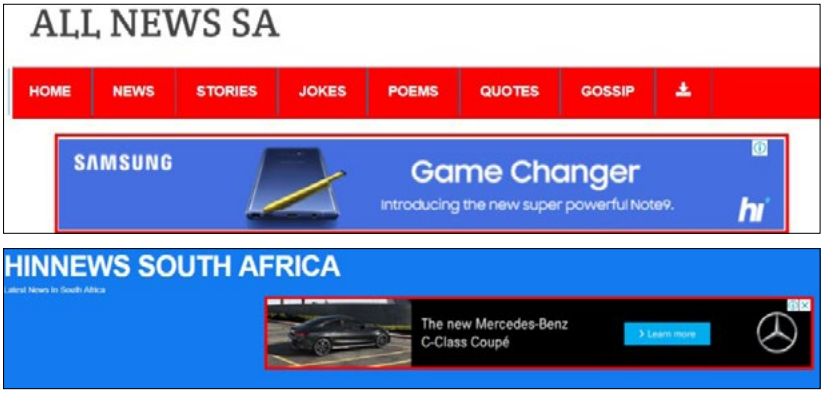


Figure 2: Samsung and Mercedes used in banner ads of fake news sites

Source: Le Roux 2018

Generative AI is increasingly being relied on to develop conspiratorial content. It is able to bypass demonetisation policies by developing content within a particular framework; for example, by focusing on cryptids and conspiracies that do not relate with violence (as conspiratorial content inciting violence is likely to be removed). However, this type of content is still harmful to society and the health of democratic processes and institutions, as it polarises and radicalises individuals while eroding trust in critical institutions and furthering 'anti-science' and conspiratorial thinking. Even clips about prehistoric megalodons roaming the ocean push the idea that the mainstream media

and mainstream social and public institutions are lying to the public – eroding trust in these institutions which becomes critical in times of disaster, like the coronavirus pandemic. These uses of AI take video clips from actual podcasts – like Joe Rogan – and insert their own voice scripts which makes it seem like Rogan himself is discussing these concepts. Owing to the history of conspiratorial conversations on Rogan's podcast in the past, combined with its massive following, it becomes a suitable target to be exploited by these content creators. This type of tactic (to use dishonest means to generate engagement with content) is known as 'engagement-farming' and brings us to our next point.

Engagement-farming

'Engagement-farming' refers to the use of various disingenuous means to artificially generate and boost engagement with content or with a particular page. These tactics can range from the use of evocative and inflammatory content that targets a particular subjective psychological vulnerability in the audience member (often identified through surveillance and analysis), to the co-opting of popular themes and discourses from other platforms or websites, and even to the imitation of popular pages or public figures (like the imitation of Joe Rogan content by generative AI mentioned previously). The ability to evade scrutiny opens the doorway to more egregious engagement-farming tactics like fear-mongering, outright disinformation, conspiracy theories, and more.

It is a popular tactic used both by content creators as well as marketers and propagandists – as well as by anti-publics like troll farms and disinformation networks. Troll farms administered by specific interests can also use imitation and conspiracies to farm engagement. For example, Myanmar's Tatmadaw – the country's military – created accounts impersonating celebrities and public figures in their disinformation operation that spread Islamophobic content and threats of violence by the Rohingya population (Mozur 2018).

Engagement-farming is successful in achieving engagement metrics and earning clickthroughs, but is detrimental to civil discourse between people. Rather than producing discourse, engagement-farming is *anti-discourse* – it brings people together on a particular issue but does so by enraging them first. This polarises people according to their emotional reactions to the content; and fractures the public sphere along these polarities.

This harms the public sphere and communicative rationality (Habermas 1989) by creating a reactive communicative environment that is volatile and emotional. This can lead to psychological states like affective override, affective polarisation, or 'amygdala hijacking' where a powerful emotional response is triggered that is capable of overwhelming typical cognitive functions – usually by presenting threatening stimuli that provokes fear-based threat responses from the audience (Gorton 2016, Iyengar et al. 2019, Mellers et al. 1992, Slovic 2007). This divides the public sphere along these emotionally entrenched lines; fracturing the sphere and harms its ability to host reasoned public discourse and exercise communicative rationality in the public sphere (Habermas 1989).

Furthermore, as this type of content triggers engagement from a user, it is subsequently personalised for that user, meaning they are likely to receive more of the same type of content. This is itself an echo-chamber-like effect, but worryingly, functions as a pipeline to more of the same content – and even more extreme forms of the same content. When a person engages with fringe content, it is read as a more personalisable interaction than when a person engages with mainstream general content. As a result, sorting algorithms will yield increasingly more fringe content for that user (Wylie 2019). This contributes to patterns of radicalisation; and often results in the audience member travelling down 'pipelines' of increasingly fringe and radical content – mediated by sorting algorithms and hosted on the platforms themselves. This brings us to the next point – patterns of migration and the effect they have on fracturing these public spheres.

The migration of content and users

Content creators seeking popular themes and discourses around which to develop engaging content has resulted in the migration of content from the fringes of the internet into the mainstream social networking platforms. These creators can even use analytical systems that help to determine which themes are popular in general, as well as which themes are popular among their particular audiences. Co-opting popular discourses and themes from more fringe websites and forums are a cheap and convenient method to make content – without the need to worry about copyright violations or other forms of IP infringement which can result in content being removed or being demonetised.

Consequently, we have seen significant amounts of traffic between the fringe and the mainstream. Popular conspiracy theories that have amassed followings on the fringe have migrated into the mainstream where they receive far more engagement and popularity. An interesting example is the facilitation of QAnon's emergence into the mainstream by wellness and lifestyle influencers on Instagram. The movement, which became known as Pastel QAnon, melded QAnon's paranoid right-wing child trafficking and anti-vaccine conspiracies with pastel colours and Instagram-esque aesthetic styles (Argentino 2021, Wilson 2020). According to some of the content creators involved, using QAnon related phrases and hashtags grew engagement with their pages by 200% (Argentino 2021, VICE 2021). The hashtag '#SavetheChildren' became a popular theme among these creators – but led to massive challenges for the actual Save The Children foundation, a UK-based anti-child-trafficking operation that became inundated with calls and claims regarding child trafficking conspiracies.

Furthermore, migration patterns have also been observed in the co-opting of cultural concerns *between* groups and communities. For example, the narratives of farm murders in South Africa were co-opted into the White Genocide and Great Replacement discourses popular among American and European right-wing audiences, who believe that Western civilisation and the 'White race' are under threat of both erasure by growing multiculturalism and historical replacement by the relative decline in White birth rates. Content creators like the Canadian Lauren Southern appropriated South Africa's farm murders to sell content to an American and European audience curated in the rhetoric of replacement theory and White Genocide (Southern 2018, CABC 2020).

This has resulted in South Africa's farm murders becoming a massive political concern among the American far-right, resulting in Donald Trump's recent critical stance towards the country (likely also the result of South Africa's actions against Israel and its involvement with BRICS). Much of this focus is based on the treatment of white citizens which is framed through the lens of the White Genocide and Great Replacement conspiracy theories as the subjugation of South African whites serves as an omen for white nationalists who see it as a precursor of their own subjugation if they become minoritised through immigration and declining birth-rates.

Crucially, when these discourses and themes from the fringe migrate to the mainstream, they serve as entry points to the fringe embedded on these mainstream networks. Users encountering these discourses in the mainstream have been observed migrating to the fringe areas of the internet in order to find more content related to these themes and discourses. These patterns of migration have been documented and studied as 'pipelines' or pathways of influence, and have been cited as a key element in the radicalisation process as users migrate down these pipelines towards echo chambers on the fringe websites and forums of the internet (Munn 2019).

The last point regarding migration is the rise of alternative platforms in response to the regulation of content. Taking action against content/pages/figures on a platform can cause users and followers to leave the platform and migrate to competitor platforms willing to host that content. This has seen the rise of many alternative platforms that operate within particular ideological frameworks – for example Gab and Truth Social which have explicitly framed themselves as conservative right-wing platforms that tolerate extremist content under claims of free speech.

These patterns of migration not only further fragment the public sphere but result in the development of new spheres that are entirely administered by the strategic interests. These interests were critiqued by Habermas as strategic rationality, which he saw as manipulating communications within the public spheres (Habermas 1989). These migratory patterns result in individuals travelling to echo chambers and spheres that are administered by increasingly secular and fringe ideologies – from 4chan and 8chan, to GAB, Rumble, and TruthSocial.

Troll farms and misinformation networks

Worryingly, these networks platform fake accounts, bots, troll farms and disinformation networks alongside real users. These fake and bot accounts provide centrally developed content disguised as individual users for the platform, and are able to influence and solicit engagement from actual users – a dynamic which benefits both the platforms and the propagandists.

The Bell Pottinger/Guptabot campaign saw South Africa's online public spheres being invaded by fake accounts that coordinated to form a disinformation network supported by a complete informational infrastructure

ranging from TV news channels to newspapers and public figures (CABC 2021). This 'botnet' was also divided up into author accounts that developed content, and amplifier accounts that circulated and boosted the content's reach. Even after Twitter attempted to purge these bots, many were still active for years. Four accounts that were identified as part of the RET disinformation network in 2016 were still active years later, and were identified as key drivers of the violent rhetoric during the looting of KwaZulu Natal during July 2021 (CABC 2021).

The Guptabot network made use of fake accounts from services like CNET Infosystems based in India (CABC 2021). This is only one such example; there are a multitude of others, forming a complete black market for disinformation, with a range of services and varying modes of monetisation and incentivisation (Gu, Kropotov, Yarochkin 2017). Services like SMOSmart, Dr Followers and CoolSouk are easily accessible services through which one may purchase likes, comments, or followers (Gu et al. 2017).

Software packages like "*Best Social Network Automation*" allow for the creation of a network of accounts along with the development of 'scripts' for the botnet to follow, allowing a user to create their own botnet. Similarly, websites like latestautomationbots.com provide these types of software packages and services for users – including account generators for the various platforms (Gu et al. 2017). These are essentially DIY propaganda kits.

Crowdsourcing has become a new model for some service providers, like VTope, which uses a CPA model (Cost Per Action). This model rewards users with digital currency for performing actions required by other users on the service – for example, submitting content takedown requests, or providing likes or comments for a page or post (Gu et al. 2017). This type of crowdsourcing services turns ordinary users into paid troll accounts. ACT-IL, a type of social network that works around conducting information operations, uses the same crowdsourcing model under the name of 'activism on behalf of the state of Israel' (Act-Il 2019). It is furnished with a 'facts library' to arm its users to engage in debates with other users online and advance a directed political ideology while masquerading as public debate between individual users (Act-Il 2019).

AIMS, a more recent technology uncovered by investigative journalists, offers more than just fake bot accounts, but complete digital avatars with interconnected accounts across Facebook and Instagram, along with accounts

on LinkedIn, AirBnb, and even credit card histories (Andrzejewski 2023). The idea is to present a very credible and realistic construction of an individual, to be wielded in information operations online.

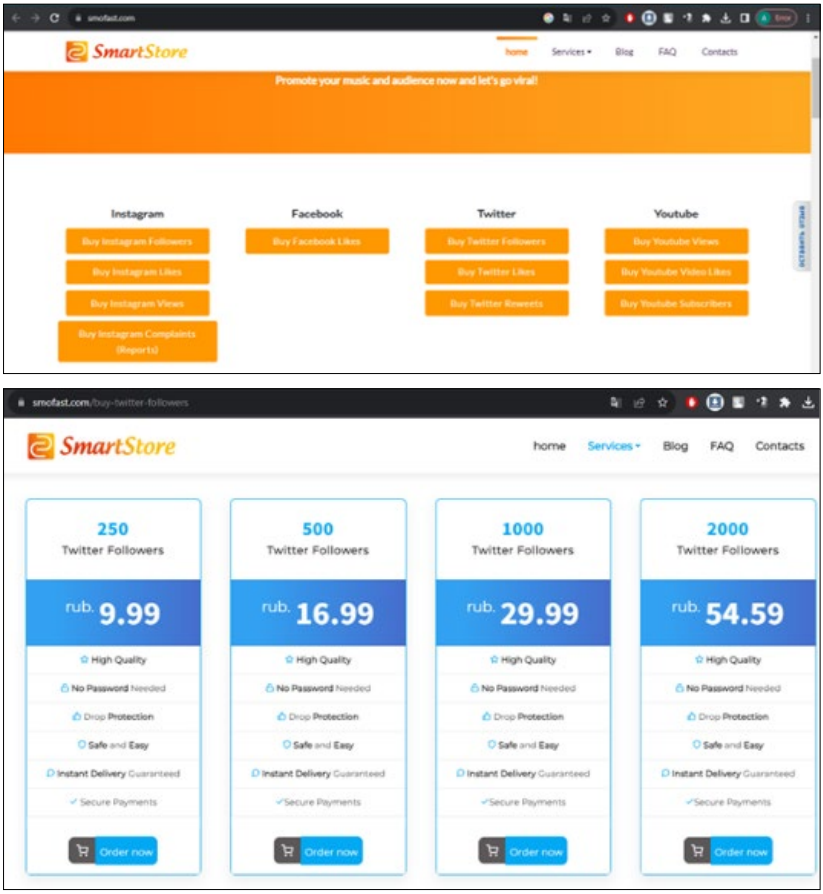
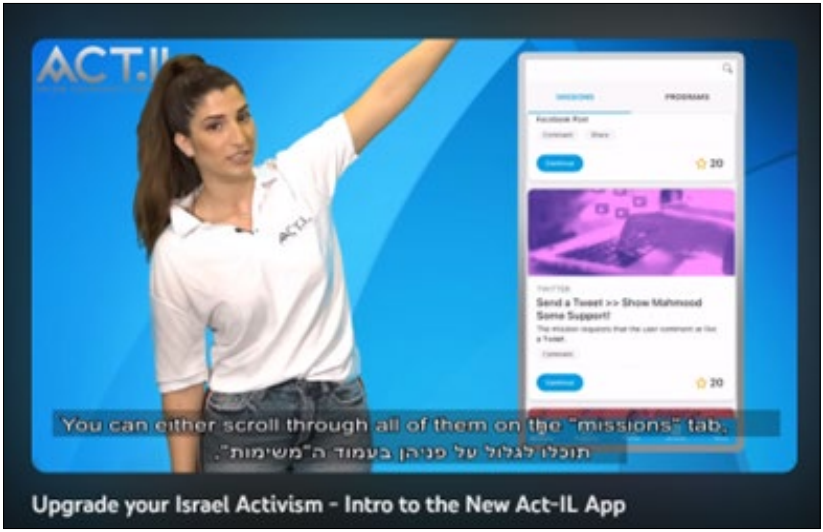
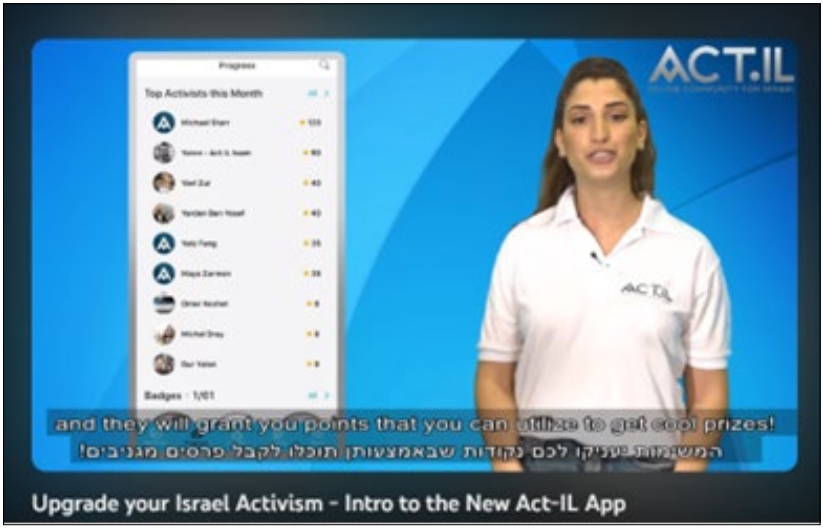


Figure 3: Screenshots from the SMOFast website indicating price ranges of various services

Source: Gu et al. 2017


ACT.IL
You can either scroll through all of them on the "missions" tab.
תוכלו לגלול על פניהן בעמוד ה"משימות".
Upgrade your Israel Activism - Intro to the New Act-IL App


ACT.IL
Top Activists this Month
and they will grant you points that you can utilize to get cool prizes!
המשימות יעניקו לכם נקודות שבאמצעותן תוכלו לקבל פרסים מגניבים!
Upgrade your Israel Activism - Intro to the New Act-IL App

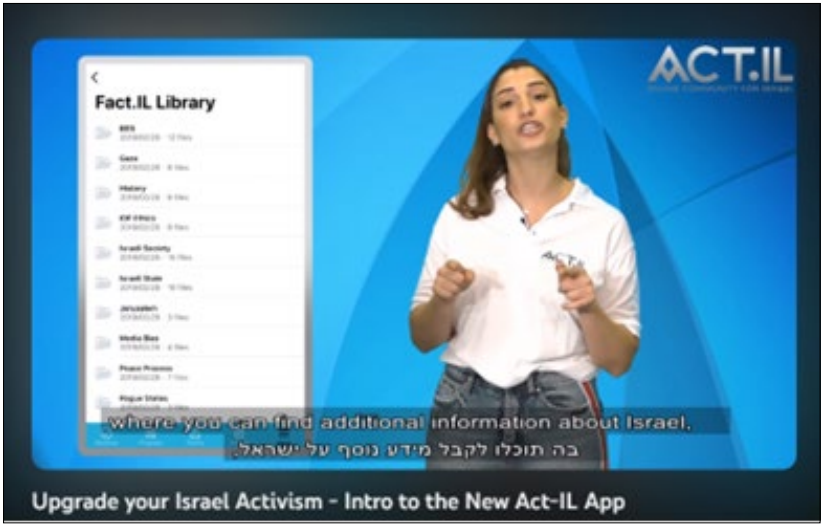


Figure 4: Screenshots from ACT-Il's YouTube channel, demonstrating the various capabilities of the social network/troll farm

Source: Act-Il 2019

Final thoughts

Habermas's concerns over the public sphere saw it as being manipulated by corporate interests that turned newspaper media into advertising media. Here the act of communication is itself undermined by 'strategic interests' stifling communicative rationality between people. However, the digital spheres are a realisation of the most extremes of these concerns; instead of being manipulated by strategic interests, these spheres are entirely administered by these strategic interests whose economic model is based on mapping and manipulating each individual user. These spheres continuously surveil and measure the responses of users in order to develop customised messaging capable of manipulating the individual user.

Furthermore, an individual is typically aware of the strategic interests at play in a piece of overt advertising content – but they are significantly less aware of the strategic interests at play in the way their social media feeds are organised, the type of content they encounter, the manner in which their own

subjectivities shape the content they receive. But as their engagements shape this content, it is analysed and used to direct individually targeted content back into their feeds, creating a self-reinforcing spiral. Combined with the tendency of the algorithms to weight highly specific engagements, this creates pull towards more and more niche and extremist content, fracturing social cohesion into increasingly isolated and polarised groups.

Today's online individual is surrounded by a 'sphere' of devices, screens and gadgets that seek to render even the most obscure or benign behaviour into usable market insights. The production and transmission of knowledge and discourse online has produced a proliferation of new subject roles that users and audience members identify with and assume. These discourses are able to herald and interpellate individuals into subject positions under discourse more effectively with the added power of automated psychological analysis conducted through surveillance of social media engagement.

For example, the anxiety around child-trafficking, combined with the availability of alarming assertions through conspiracy theories like QAnon, saw many individuals assume the role of concerned citizen-activist. This disrupted actual anti-child trafficking agencies who were inundated with conspiratorial claims. Similarly, US voter registry departments have had to filter out thousands of claims about fraudulent voters that have been put forth by several individuals acting in the name of citizen-journalism. South African police had to respond to false child trafficking claims as QAnon's online following extended to South Africa. The uncertainty and frustration during the Covid pandemic resulted in concerned users sharing dangerous misinformation – from anti-science discourse that rejected the messaging of critical public health institutions to home-made remedies often involving experimenting with dangerous substances and treatments. These polarised subjectivities not only undermine Habermas's aspirations to communicative rationality and a democratic public sphere, but threaten the social institutions that protect people, such as the public health and law enforcement organisations detailed above.

We thus propose developing a formalised Informational Vulnerability Index that assesses the linguistic, cultural, and historical history of a region in order to better govern the operations of social networks in that region. For example, a country like South Africa, with a history of Apartheid, as well as with a diverse cultural mix of peoples and languages, would require a diverse team of moderators owing to the various dialects spoken – as well as stricter

controls over advertising practices and the sorting of content. This index seeks to categorise and analyse radical and conspiratorial discourses as well; making note of the *availability* of corresponding information, the *accessibility* of that information.

Informally, we suggest championing a culture of ethical podcasting which not only steers away from engagement-farming, but recognises the harms not only in disinformation or misinformation, but the harm of incomplete information without context. Narratives like 'just asking questions' are used to legitimate hosting provocative discussions that generate engagement and curiosity while steering the audience away from traditional media towards these new media forms. These new media forms have amassed such large followings that it has ironically become the new mainstream media.

This intervention and regulation is essential to protect the communicative rationality and the public sphere against the harms being produced by both deliberate political and economic manipulation through individual psychographic targeting of social media feeds, and the harms produced by the polarising and radicalising drift of media feeds designed to maximise engagement as core of the business model of social media platforms. Against this business model, a system that disincentives anti-democratic and socially harmful content and systems of distribution would need to be developed, while avoiding the political risks of centralised media control and censorship.

As we have thus explored, the arrival of the internet and social networking has not delivered on the promise of a potentially democratic public sphere. Habermas's vision of a public sphere as an arena for communicative rationality made available to all has not been realised. Instead, what has been realised is an even more insidious version of Habermas's concerns about strategic interests manipulating the public sphere. New technologies allow these digital public spheres to not only be manipulated by strategic interests, but to be *administered* by the strategic interests themselves. Google, Meta, X, and all the other online platforms and services are governed directly by the strategic rationality of the day – Surveillance Capitalism, leaving behind a vacuum that is filled with manipulative content, misinformation, and anti-publics masquerading as information, communication and community.

Our work is part of rallying cry to reclaim these spaces that have been colonised by Surveillance Capitalism, spaces that could potentially harbour egalitarian discourse and communicative rationality. Not only did Surveillance

Capitalism claim the digital world and the potential of the digital public sphere, but it used our connection to the digital world to claim our personal lives and personal experiences as resources – an untapped virgin wood to process into data and render into market insights and technologies of covert manipulation.

References

Act-IL. 2019. Upgrade your Israel activism - intro to the new Act-IL app. Available at: https://www.youtube.com/watch?v=... [accessed on 13 November 2025].

Ahmed I. 2023. Toxic Twitter: how Twitter generates millions in ad revenue by bringing back banned accounts. *Centre for Countering Digital Hate*. 9 February. Available at: https://counterhate.com [accessed on 13 November 2025].

Amiradakis MJ. 2019. Habermas, mass communication technology and the future of the public sphere. *South African Journal of Philosophy* 38(2): 149–165. https://doi.org/10.1080/02580136.2019.1620515

Andrzejewski C. 2023. "Team Jorge": in the heart of a global disinformation machine. *Forbidden Stories*. 15 February. Translated from French by A Hylton. Available at: https://forbiddenstories.org [accessed on 13 November 2025].

Argentino MA. 2021. Pastel QAnon. *Global Network on Extremism and Technology*. 17 March. Available at: https://gnet-research.org [accessed on 13 November 2025].

Centre for Analytics and Behavioural Change. 2020. Farmlands - the making of a misleading 'documentary'. *Centre for Analytics and Behavioural Change*. 8 December. Available at: https://iono.fm [accessed on 13 November 2025].

Centre for Analytics and Behavioural Change. 2021. Online RET network analysis. *Centre for Analytics and Behavioural Change*. Available at: https://cabc.org.za [accessed on 13 November 2025].

Centre for Analytics and Behavioural Change. 2021. The dirty dozen and the amplification of incendiary content during the outbreak of unrest in South Africa July 2021. *Centre for Analytics and Behavioural Change*. Available at: https://drive.google.com/file/d/1TZV8ZfX1oAXWqFpeGth_fqBSNse2JrxR/view?usp=drive_link [accessed on 13 November 2025].

Dahlgren P. 2005. The Internet, public spheres, and political communication: dispersion and deliberation. *Political Communication* 22(2): 147–162. https://doi.org/10.1080/10584600590933160

Golin J. 2015. Child advocates mobilize to stop Mattel's eavesdropping "Hello Barbie". *Fairplay: Childhood Beyond Brands*. 10 March. Available at: https://fairplayforkids.org [accessed on 13 November 2025].

Gorton W. 2016. How microtargeting harms the public sphere. *YouTube*. Available at: https://www.youtube.com/watch?v=... [accessed on 13 November 2025].

Gu L, Kropotov V and Yarochkin F. 2017. *The fake news machine: how propagandists abuse the internet and manipulate the public*. Trend Micro. Available at: https://trendmicro.com [accessed on 13 November 2025].

Habermas J. [1962] 1989. *The structural transformation of the public sphere: an inquiry into a category of bourgeois society*. Translated by T Burger. Cambridge: MIT Press.

Iyengar S, Lelkes Y, Levendusky M, Malhotra N and Westwood SJ. 2019. The origins and consequences of affective polarization in the United States. *Annual Review of Political Science* 22: 129-146. https://doi.org/10.1146/annurev-polisci-051117-073034

Le Roux J. 2018. Programmatic advertising. *Exposed News24*. October. Available at: https://news24.com [accessed on 13 November 2025].

Le Roux J. 2018. EXPOSED: the Unisa employee who manufactures fake news to divide SA. *Exposed News24*. 15 November. Available at: https://news24.com [accessed on 4 August 2022].

Mellers BA, Schwartz A, Ho K and Ritov I. 1997. Decision affect theory: emotional reactions to the outcomes of risky options. *Psychological Science* 8(6): 423-429. https://doi.org/10.1111/j.1467-9280.1997.tb00455.x

Mozur P. 2018. A genocide incited on Facebook, with posts from Myanmar's military. *New York Times*. 15 October. Available at: https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html [accessed on 13 November 2025].

Munn L. 2019. Alt-right pipeline: individual journeys to extremism online. *First Monday* 24(6). https://doi.org/10.5210/fm.v24i6.10108

Mylonas Y. 2023. On the proletarian public sphere and its contemporaneity: crises, class and the media. *Acta Academica* 55(2): 111–130. https://doi.org/10.38140/aa.v55i2.7728

Pariser E. 2011. *The filter bubble: what the internet is hiding from you*. United Kingdom: Penguin. https://doi.org/10.3139/9783446431164

Reuters. 2024. Ex-Meta employee claims firing was due to Gaza content handling. *The Jerusalem Post*. 5 June. Available at: https://www.jpost.com/israel-hamas-war/article-805046 [accessed on 13 November 2025].

Slovic P, Finucane MP, Peters E and MacGregor DG. 2007. The affect heuristic. *European Journal of Operational Research* 177(3): 1333-1352. https://doi.org/10.1016/j.ejor.2005.04.006

Southern L. 2018. *Farmlands* [documentary film]. 25 June. Available at: https://youtube.com [accessed on 13 November 2025].

Strydom A. 2017. *Manufacturing divides: the Gupta-linked radical economic transformation (RET) media network*. African Network of Centres for Investigative Reporting. Available at: https://sourceafrica.net [accessed on 13 November 2025].

Thiel P. 2009. The education of a libertarian. *Cato Unbound*. 13 April. Available at: https://www.cato-unbound.org/2009/04/13/peter-thiel/the-education-of-a-libertarian/ [accessed on 13 November 2025].

United States of America. 104th United States Congress. 1996. *Section 23 Communications Decency Act of 1996. Protection for 'good samaritan' blocking and screening of offensive material.*

VICE News. 2021. This lifestyle influencer lost her accounts because of QAnon. *VICE News*. 22 February. Available at: https://youtube.com [accessed on 13 November 2025].

Wilson S. 2020. The wellness realm has fallen into conspiritualism – I have a sense why. *The Guardian*. 14 September. Available at: https://www.theguardian.com [accessed on 13 November 2025].

Wylie C. 2019. *Mindf*ck: inside Cambridge Analytica's plot to break the world*. London: Profile Books.

Zuboff S. 2019. *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. London: Profile Books.