Funded by the European Union

SMIDGE

Social Media Narratives : Addressing Extremism in Middle Age

D4.5 Report on Networking Technologies (update)

[27 February 2026]

DE MONTFORT UNIVERSITY LEICESTER
FAS research
CENTRO STUDI VILLA MONTESCA FONDAZIONE
INCLUSIVE EUROPE
KCSS
University of Central Lancashire UCLan Cyprus
KØBENHAVNS UNIVERSITET
UNIVERSITÀ DEGLI STUDI DI MILANO

Disclaimer

This project has received funding from the European Union's Horizon Research and Innovation Programme Under Grant Agreement no 101095290.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

More info and contact: info@smidgeproject.eu | www.smidgeproject.eu

Funded by the
European Union

Document Control

Deliverable	D4.5
WP/Task Related	WP4/T4.3
Delivery Date	27/02/26
Dissemination Level	PU – Public
Lead Author	Raouf Hamzaoui
Lead Partner	DMU
Contributors	
Reviewers	Line Nybro Petersen (UCPH), Ramadan Ilazi (KCSS)
Executive Summary	Social networking platforms have fostered communication and community building but also provided extremists with tools to spread their ideologies, coordinate attacks, and publicise violence. High-profile incidents, such as the Halle attack livestreamed on Twitch, the Christchurch mosques shooting livestreamed on Facebook, and the storming of the US Capitol livestreamed on DLive, illustrate the dangers of extremists exploiting these platforms. Fringe platforms, with minimal moderation, have become hotspots for extremist activity, attracting users with promises of "free speech". Mainstream platforms, despite more regulation, also contribute to radicalisation, with algorithms potentially pushing users towards extremist content. Extremists are increasingly using advanced technologies like deepfakes and AI chatbots to manipulate public opinion and create extremist content. These developments underscore the urgent need for effective governance, transparency, and enforcement mechanisms. This report emphasises the role of advanced machine learning, particularly deep learning, in the creation of extremist content. Many platforms allow pseudonymous registration, masking user identities but not fully preventing tracking via IP addresses, emails, or phone numbers. End-to-end encryption, increasingly used for messaging and video calls, provides some privacy, which can be enhanced with virtual private networks. However, most platforms rely on centralised servers, creating significant privacy risks. Tor and Lokinet can improve anonymity by routing traffic through relay networks. The business models of these platforms, driven by advertising, donations, crowdfunding, and premium subscriptions, affect content management and may encourage tolerance of extremist content. Extremists exploit the ability to monetise content through tips, donations, and cryptocurrencies, offering additional funding avenues for their activities. Deepfakes, created using deep neural networks, pose growing threats by enabling disinformation and incitement to violence. Although current deepfakes often suffer from low quality, rapid technological advancements may soon increase their impact. Deepfake generation

	methods include autoregressive models, autoencoders, generative adversarial networks, and diffusion models, which are becoming increasingly accessible through free and simple to use applications. Deepfake detection methods face limits in robustness and generalisation, especially across datasets, generator families, and real-world conditions. Extremists are also using bots to spread propaganda, share expertise, and manipulate social media metrics, creating the illusion of widespread support for their content. Freely available generative chatbots based on large language models can be exploited by extremists to generate extremist content. They can also provide expertise that is otherwise hard to obtain. For example, tools like ChatGPT and Microsoft Copilot can provide instructions and computer code for evading online detection, as well as information on violent tactics and weaponry. Emerging, still-maturing technologies could pose a near-future threat if adopted by far-right actors for propaganda and mobilisation. Of particular concern are audio-visual deepfakes, synthetic identities, hyper-personalised targeting, affective computing, and VR/AR immersive influence. This report is organised as follows. Section 2 provides a comparative analysis of these platforms based on six criteria: launch year, core functions, networking technology, privacy and security, business model, and content moderation. Section 3 examines deepfake technology and its use by extremists, including deepfake generation and detection. Section 4 is dedicated to bots. Section 5 examines emerging technologies. Section 6 summarises the findings.
Key Words	Blockchains, bots, content filtering, deepfakes, deep learning, far-right, peer-to-peer networks, social media networks.

Revision History

Version	Date	Author(s)	Reviewer(s)	Notes
0.1	16/02/26	Raouf Hamzaoui	Line Nybro Petersen (UCPH), Ramadan Ilazi (KCSS)	
0.2	27/02/26	Raouf Hamzaoui		

List of Acronyms/abbreviations

Abbreviation	Full Name
Acc	Accuracy
AP	Average Precision
AR	Augmented Reality
AUC	Area Under Curve
CDN	Content Delivery Network
CNN	Convolutional Neural Network
DDoS	Distributed Denial of Service
DHT	Distributed Hash Table
DNN	Deep Neural Network
DSA	Digital Services Act
E2EE	End-to-end Encryption
GAN	Generative Adversarial Network
HTTPS	Hypertext Transfer Protocol Secure
IP	Internet Protocol
LLARP	Low Latency Anonymous Routing Protocol
LLM	Large Language Model
NLP	Natural Language Processing
P2P	Peer-to-Peer

ResNet	Residual Network
RNN	Recurrent Neural Network
TCP	Transmission Control Protocol
TLS	Transport Layer Security
Tor	The Onion Router
UDP	User Datagram Protocol
VAE	Variational Autoencoder
VPN	Virtual Private Network
VPS	Virtual Private Server
VR	Virtual Reality

Table 1: List of acronyms/abbreviations

Table of Contents

1 Introduction

The proliferation of social networking platforms has created unprecedented opportunities for communication and community building. However, this digital landscape has also been exploited by extremists to disseminate their ideologies, coordinate actions, and publicise their attacks. The use of social networking platforms by extremists is a growing concern, underscored by several high-profile events in recent years [Kupp22].

On 9 October 2019, far-right extremist Stephan Balliet uploaded a manifesto to the Meguca message board a few minutes before attacking a synagogue in Halle and streaming the attack live on Twitch [Koeh19]. The perpetrator of the 2018 Pittsburgh terrorist attack posted an antisemitic message on Gab before killing 11 people and injuring six [Roos18]. The perpetrator of the 2019 shooting at a mosque in Bærum posted messages on Endchan shortly before the attack [Hume19]. A few minutes before the mass shooting in two Christchurch mosques on 15 March 2019, Brenton Tarrant announced his intent to carry out an attack on 8chan. He livestreamed the massacre on Facebook using a GoPro camera. Around 200 people watched the live broadcast. None of them reported the incident to Facebook. Within the first 24 hours, around 1.5 million copies of the video were removed from Facebook, while tens of thousands were deleted from YouTube [Mack19]. A recording of the mass shooting was also uploaded to BitChute on the same day, accumulating over 190,000 views by 21 March [Thom20]. On 3 August 2019, Patrick Crusius killed 23 people and injured 22 others at a Walmart store in El Paso. Shortly before the shooting, he posted a manifesto on the imageboard 8chan [Conr20]. The manifesto cited the Great Replacement[1] conspiracy theory as one justification for the massacre [Evan19]. Although site moderators quickly removed the original thread, anonymous users continued to post links to the manifesto [Evan19]. The storming of the US Capitol on 6 January 2021 was streamed live on DLive [Brow21a]. On 14 May 2022, Payton Gendron carried out a racially motivated mass shooting at a supermarket in Buffalo, killing 10 people and injuring three. He had an online diary on Discord which included plans for the attack and showed that he was radicalised by reading online about the Great Replacement conspiracy theory and inspired by other far-right mass shootings. He livestreamed the attack on Twitch using a GoPro camera [Brom22].

In addition to these incidents, extremists are increasingly using sophisticated technologies such as deepfakes and bots to achieve their goals. Deepfakes, which are AI-generated synthetic media, allow extremists to create convincing fake audio, images, and video, influencing public opinion and inciting violence. In 2023, far-right activist Jack Posobeic posted a deepfake video of President Biden announcing the return of the military draft for Americans due to growing tensions with Russia on Ukraine [Thal23]. Chatbots based on large language models (LLM) such as Open AI's ChatGPT, Meta's LLaMA, Microsoft Copilot, and Google's BERT and Gemini can be used to quickly generate extremist content [Bael22]. Other bots are used to disseminate extremist content and disrupt social media platforms.

This report focuses on fringe platforms, which, unlike mainstream platforms[2], are less regulated and consequently more susceptible to misuse by extremist groups. Owners of fringe platforms often promote their sites as "free speech" alternatives to mainstream platforms. This promise of limited moderation has attracted extremists. While mainstream platforms are subject to more regulation, they have also contributed to user radicalisation and recruitment of extremists. A large-scale study [Ribe20] of comments and recommendations of three YouTube communities suggests that users progress towards more extreme content on the platform. YouTube's recommendation algorithm is designed to maximise user engagement and advertising revenue. This algorithm, based on machine learning through deep neural networks [Covi16], may have inadvertently discovered a correlation between racism and prolonged viewer engagement, which would explain a reported bias toward suggesting alt-right content to users [Brya20].

By examining current use of these platforms, along with the advanced techniques used by extremists, this report aims to shed light on the evolving landscape of extremist activities online and the critical need for

[1] The Great Replacement is a far-right conspiracy theory that states that Western elites are deliberately replacing white people living in Western countries with non-white immigrants.

[2] A mainstream platform is a widely used, well-known, and easily accessible platform.

effective monitoring and regulation. The report highlights the growing importance of advanced machine learning, specifically deep learning, in the creation, detection, and filtering of extremist content. Deep learning uses deep neural networks, artificial neural networks consisting of multiple layers of nodes (neurons). Each node is connected to other nodes in the previous layer, with each connection having a weight and each node having a bias.

The report is organised as follows. Section 2 provides a comparative analysis of these platforms based on six criteria: launch year, core functions, networking technology, privacy and security, business model, and content moderation. Section 3 discusses deepfake technology and its use by extremists, including deepfake generation and detection. Section 4 is dedicated to bots. Section 5 examines emerging technologies. Section 6 summarises the findings.

2 Platforms

To disseminate hate messages, calls to violence, propaganda, radicalisation materials, recruitment messages, misinformation, disinformation, and conspiracy theories, extremists use both mainstream social networking platforms like Facebook and fringe “alt-tech” platforms.

In this section, these platforms are reviewed. Six characteristics are considered: the year the platform was launched, main functionalities, networking technology, privacy and data security, content moderation, and business model to generate income.

This review focuses on alt-tech platforms as they are less known than mainstream ones. While mainstream platforms do not intentionally promote harmful content, their algorithms favour right-leaning political videos, including racist views from the alt-right [Brya20]. This is because the aim of the algorithms is to drive users to use the service as much as possible, optimising the chance that the user will click an ad, generating revenue for the website.

Due to the large number of existing platforms and their similarities, only selected ones were reviewed. For example, there are many alt-tech imageboards with minimal moderation called Chans: 8chan, 8kun, 9chan, 16chan, InfinityChan, ShitChan, EndChan, and NeinChan. Some no longer exist or are accessible only on the dark web [Sale22] (e.g., 16chan and NeinChan). All these boards host a /pol board dedicated to "politically incorrect" discussions. 8chan was renamed 8kun after Internet security firm Cloudflare discontinued its service, making it vulnerable to distributed denial of service (DDoS) attacks. Neinchan, launched in February 2019, was initially accessible on the Internet but moved to the dark web in April 2020 when its hosting service, Clearnet, terminated it due to extremist content [Craw19]. A 2021 study of six /pol chan boards [Bael21] found that while these boards share a common far-right white supremacist subculture, this subculture has fragmented in terms of the extremity of views. This fragmentation can be described as a three-tiered hierarchy, where boards higher in the hierarchy are more popular but feature less extreme content.

Table 2 summarises the findings. When information was unavailable, the corresponding field was left blank. The data was collected from April 2023 and verified in February 2026. Given the continuous evolution of these platforms, some information may become outdated over time.

Platform	Launch	Features	Registration	Moderation	Network	Business model
Banned.video		Video hosting, livestreaming	Email	Human	Centralised	Donations, payments for products sold on Infowars Store
Bitchute	2017	Video hosting	Email	Flagging and reviewing	Centralised	Donations, advertising (restricted),

						account upgrade
Brighteon	2018	Video hosting, video streaming	Email, phone number	Human	Centralised	Creator/vendor payout, donations
Discord	2015	Voice and video call, text messaging, post messages and video, live video, gaming, screen sharing	Email	Human moderation and bots	Centralised	Subscription fee, customisation
Cozy.tv	2021	Livestreaming	Phone number	Human	Centralised	Donation
DLive	2017	Live video, live chat rooms, gaming. Streaming to the platform (Dlive.tv) or directly to viewers (Dlive)	Email	Human	Centralised (TRON network for rewards/monetisation).	Donations with blockchains, subscription fee
Endchan	2015	Post messages on board	Email	Users can moderate their own boards	Centralised (hosted on a VPS, reachable via Tor or Lokinet)	Donations
Entropy	2019	Livestreaming	Email	Human	Centralised (via Microsoft's Azure cloud)	Monetised streaming
Facebook	2004	Post messages, images, video. Send messages to other users. Video and voice calls. Livestreaming	Email/ mobile number	Human moderation, machine learning	Centralised	Targeted advertisement. Payments for games played and products sold on the platform
Gab	2016	Post messages, images, video. Messaging. Livestreaming	Email. User verification possible	Human	Centralised	Premium subscriptions, donations, affiliate partnership
Gettr	2021	Post messages, images, video. Messaging. Livestreaming	Email. Verification	Community reporting	Centralised	Venture capital funding
Kiwi Farms	2013	Web forum	Email.	Human	Centralised (via CDN)	Donations
Mastodon	2016	Open-source software for setting up self-hosted	Set up own server or join an existing	Moderation by server administrator	Decentralised (federated)	Sponsorships, crowdfunding

		social networks (servers)	one. Email. Verification			
Minds	2015	Open source social network	Mobile number to earn tokens	Human	Centralised (supports ActivityPub federation)	Premium subscription. Equity crowdfunding
Odysee	2020	Share video and other content	Email	Creator moderation	LBRY network (P2P distribution; metadata on-chain)	Credits
Parler	2018	Post text and images, video streaming	Email or phone number	Human	Centralised (via own cloud service)	Advertising
Patriots.win	2019	Forum	Email	Human	Centralised	Advertising
Reddit	2005	Forum, sharing text, images, video; chat	Email	Human	Centralised (via Amazon Web Services)	Advertising, premium subscription
Rumble	2013	Video hosting, video streaming	Email. Optional verification with mobile number	Moderation by creators and consumers. Automated flagging	Centralised (via own cloud service)	Advertising, private investment
Scored	At least since 2020	Posting messages	User name and password	Human	Centralised	Advertising, pro subscriptions, donations
Signal	2014	Instant messaging, voice calls, video calls	Phone number	N/A	Centralised	Grants, donations
SimpleX Chat	2022	Voice call, video call, instant messaging	None	Human	Decentralised	Venture capital and private investors, donations
Stormfront	1996	Post messages to forum, upload files, send messages to other members	Email	Human	Centralised	Donations
Telegram	2013	Voice/video calls, file sharing, group chats, channels	Mobile number which can be hidden to other users	Yes, but can be disabled	Centralised	Investment, advertising
The Daily Stormer	2013	Message board			Centralised (hosted as a Tor onion service)	Donations
Truth Social	2022	Posts, direct messaging	Email. Optional verification	Human moderation and machine learning filtering	Centralised (via CDN)	Private investors
Twitch	2011	Livestreaming and on demand, especially	Email or mobile number	Human and automated via machine learning and natural	Centralised (via AWS cloud services)	Advertising, donations, payments for extra features

		gaming, direct messaging		language processing algorithms		
Twitter/X	2006	Post text, images, video. Private messages	Email or mobile number. Optional verification	Human moderators and AI	Centralised	Advertising, data licensing, subscription
YouTube	2005	Video sharing, livestreaming	Google account	Human moderators, machine learning, community reporting	Centralised	Advertising, premium subscription
4chan	2003	Post images, text	Account not required. Tripcodes optional for authentication	Human	Centralised	Advertising, donations
8kun	2013	Post images, text. Send private messages to other accounts. Does not hold or store old content permanently and old threads are pruned as new ones are created.	Anonymous posting	Human	Centralised	Subscription

Table 2: List of platforms

2.1 Launch

The reviewed platforms, shown in Table 2, were launched between 2003 and 2022. The oldest one is the imageboard 4chan, and the most recent one is Truth Social, which was launched by Donald Trump in 2021 after he was banned from Twitter and Facebook. All listed platforms were active on 26 February 2026. Parler was removed from Apple's and Google's app stores in January 2021 due to inadequate moderation of posts that encouraged violence following the 6 January 2021 U.S. Capitol riot [Parl], [PeLy21]. It then went offline later that month after Amazon Web Services (AWS) terminated its hosting [Hern21]. Parler went offline again in April 2023. It started coming back in March 2024, returning to Apple's App Store as an invite-only app [KeTu24].

Figures 1-6 illustrate the content of several reviewed platforms.

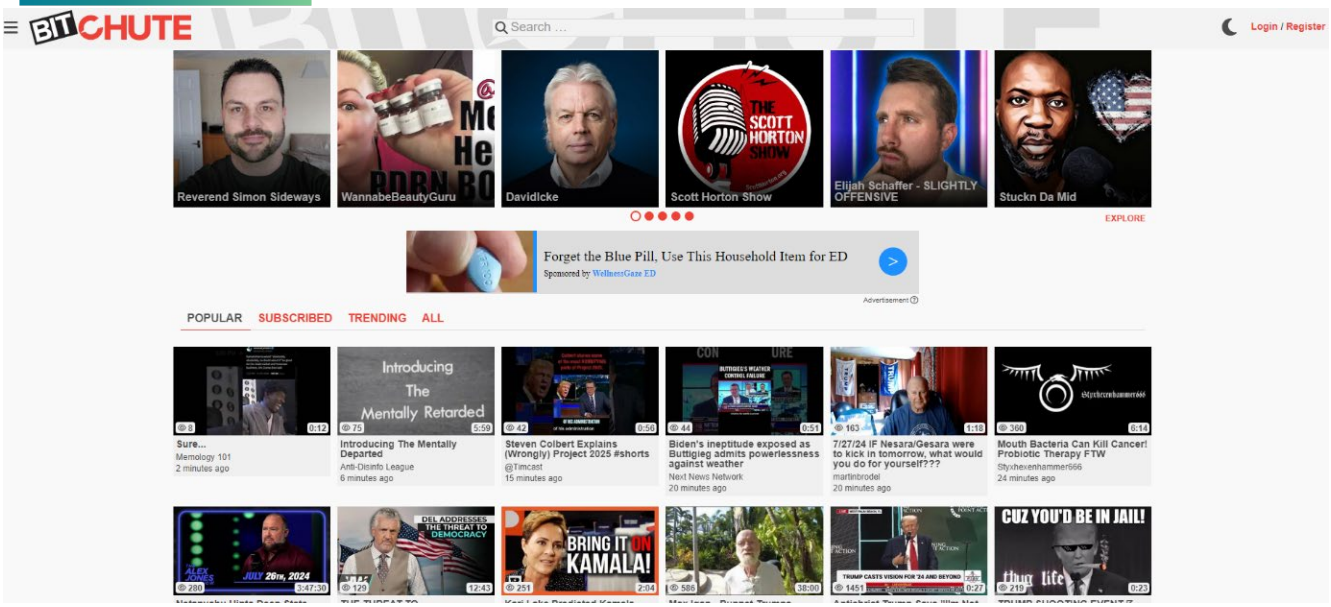

Fig. 1: Bitchute

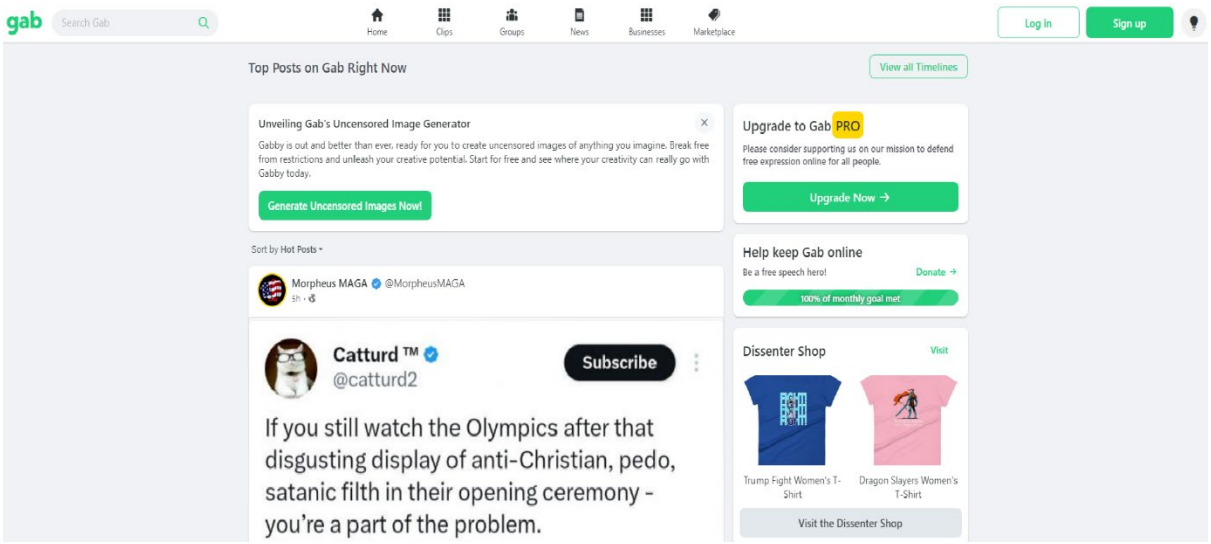

Fig. 2: Gab

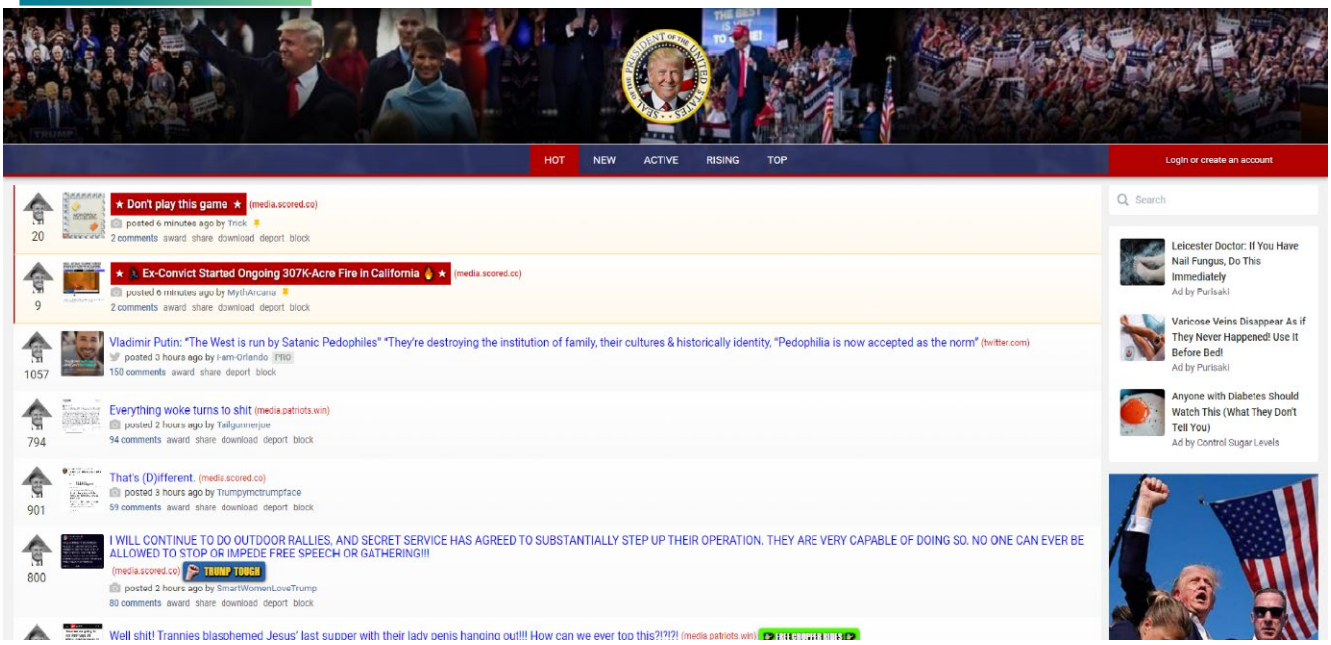

Fig. 3: Patriots.win

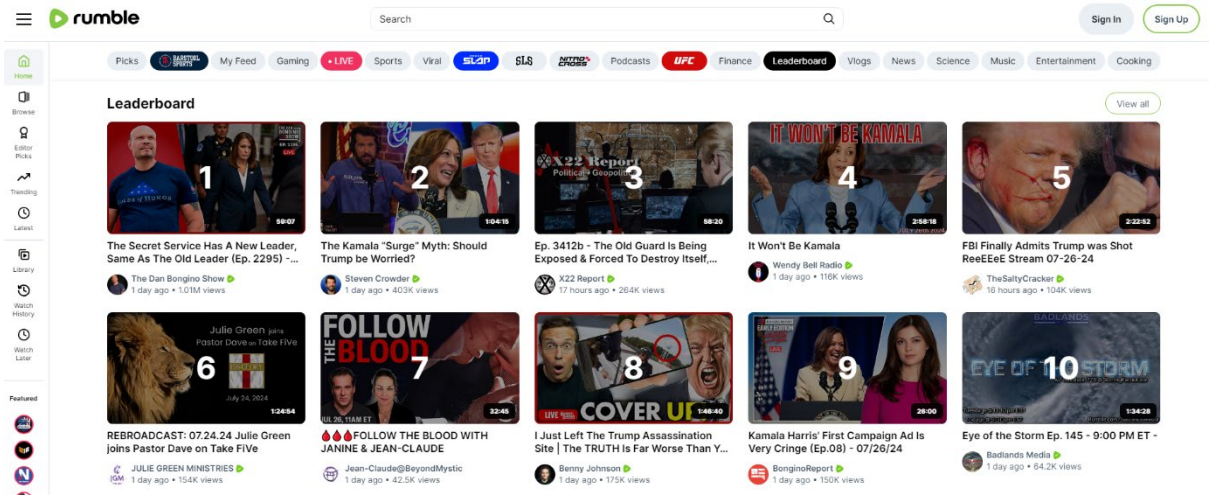

Fig. 4: Rumble

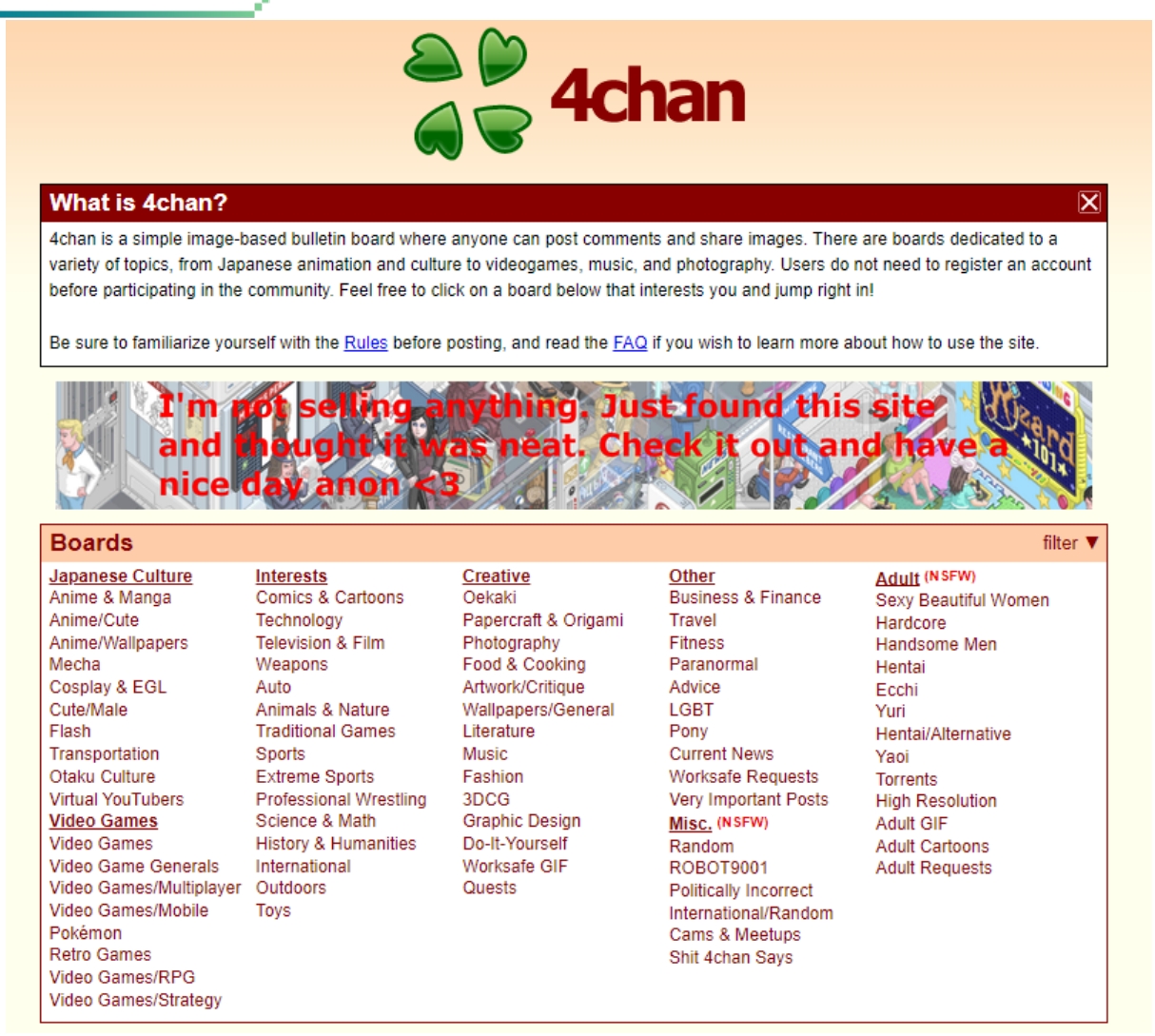

Fig. 5: 4chan

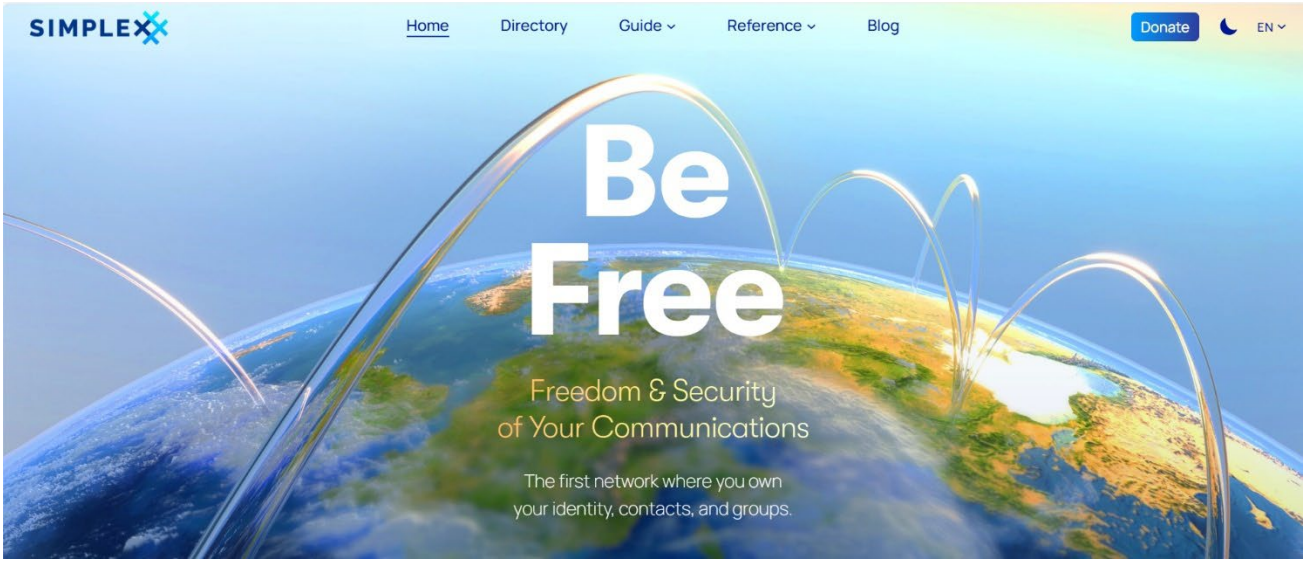

Fig. 6: SimpleX Chat

2.2 Features

Social networking platforms offer a wide range of functionalities, including posting text, files, audio, images, and video on the platform; making audio and video calls; video conferencing; live video streaming; screen capture streaming for, for example, streaming video games live. Many platforms allow users to create groups to restrict communication to the group members. Users can send private messages to other users (direct messaging) or share files with groups of users. Some platforms (e.g., Endchan) allow users to create their own boards. Other functionalities include editing video, reacting to content (e.g., Facebook's Like, Love, Haha, Wow, Sad, Angry), sharing other users' posts on one's profile (e.g., Twitter's retweets), importing posts from other platforms, and making donations to content creators or to the platform. In addition, users can use social media management tools (e.g., Buffer or Hootsuite) to post the same media content on several social media platforms simultaneously (cross-posting). This feature allows users to save time and reach a wider audience.

The platforms can be classified according to the following categories:

- **Microblogging**: users can post short sentences, images, and video links. Examples include Gab, Gettr, Facebook, Parler, Truth Social, X/Twitter.
- **Internet forum (message board):** users can post text to discuss topics. Examples include The Daily Stormer, Kiwi Farms, Patriots.win, Reddit.
- **Imageboards**: users can post images and discuss topics. These platforms are like message boards, with a focus on visual content. Examples include 4chan, Endchan, 8kun.
- **Video hosting**: users can upload and share videos. On some platforms, users can also livestream videos. Examples include Bitchute, Odysee, Rumble, and YouTube.
- **Livestreaming**: these platforms focus on streaming live video. The video can be captured with a camera or from the computer screen. Examples include DLive, Entropy, and Twitch.
- **Instant messaging**: users can communicate with other users in real-time using text, voice, and video. Communication can be private or take place in chat rooms, which can be accessed via links. Examples include Discord, Signal, Telegram.
- **Software**: these platforms offer a backend to other platforms for setting up self-hosted social networks. Examples include Minds and the open-source microblogging software Mastodon.

The distinction between these categories can be fluid, as platforms may evolve and add new features. For example, X/Twitter and Gettr started as microblogging platforms before introducing livestreaming.

2.3 Networks

Most platforms are centrally managed by a single provider and hosted on dedicated or cloud infrastructure, such as Amazon Web Services or Microsoft Azure. Servers may be located in a single region or distributed across multiple data centres, often supported by Content Delivery Networks (CDNs). Some platforms choose to host servers within a single jurisdiction to reduce exposure to multiple legal systems.

An alternative to centralised networks is decentralised networks. Examples include The Onion Router (Tor), Invisible Internet Project (I2P), and the Low Latency Anonymous Routing Protocol (LLARP). When nodes act as both clients and servers, the network is called peer-to-peer (P2P).

Tor is free software that enables anonymous communication over the Internet using a distributed network of relay nodes run by volunteers. The network consists of three types of relay nodes: guard nodes, middle nodes, and exit nodes. At least three such relays are required to build a connection. The client establishes a separate encryption key with each relay, and each relay removes one layer of encryption before forwarding the traffic. Users' data is routed through these relay nodes, concealing the IP addresses of the sender and receiver. Tor is

not entirely decentralised, as it relies on central servers called directory authorities that maintain information about the state of the network. The directory authorities are operated by a small group of trusted operators. Tor natively supports TCP streams but not UDP traffic, which limits certain applications. Because of the limited bandwidth of relay nodes and multiple layers of encryption, end-to-end latency can be high and data transfer rates can be low [Azad23]. Tor is also susceptible to Sybil attacks, in which an adversary creates many malicious nodes. An attacker who controls a sufficient number of relays may increase the probability of compromising anonymity [Tosi22].

I2P is a P2P network in which participants operate routers that forward traffic for one another. I2P uses garlic routing protocol tunnels, address books, and a distributed network database [Sale22]. Unlike Tor, I2P uses a Distributed Hash Table (DHT) for peer discovery and routing. Unlike Tor, I2P supports both TCP and UDP traffic [Tosi22]. However, like Tor, I2P is susceptible to Sybil attacks [Tosi22].

LLARP aims to provide faster and more secure communication than Tor and I2P. LLARP does not rely on central directory authorities and instead uses a DHT. The list of service nodes is derived from blockchain staking transactions, and the DHT operates over these nodes [Tosi22], [LLAR]. A blockchain is a distributed database that maintains a continuously growing list of records, called blocks, which are linked together using cryptography. Lokinet is an implementation of LLARP. Lokinet uses Oxen service nodes as relays and a cryptocurrency-based incentive system to encourage node operation. Lokinet stores the state of the service node network on the blockchain, and participating nodes reach consensus over this list. This design removes the need for central directory authorities as used in Tor.

In addition to hosting its main website on centralised server infrastructure, Endchan can also be accessed through decentralised anonymity networks such as Tor and Lokinet. According to Endchan's FAQ, the site is reachable via a Tor hidden service and has a Lokinet address, which allow users to connect without exposing their IP addresses or physical location [Endchanfaq].

Odysee is a video platform built on the LBRY protocol (https://lbry.com/). LBRY uses a blockchain to store content metadata and "claims" (content entries), while the media files are distributed through a P2P network. The Odysee website and apps, however, are centrally operated interfaces that can choose what to show or block.

DLive is a centrally operated streaming platform that integrated with the TRON network (https://tron.network/) after joining BitTorrent. It uses blockchain mainly for rewards and monetisation and has linked its storage plans to BitTorrent File System (BTFS), but the platform's core service and governance remain centralised.

Mastodon is a federated social network made up of independent servers (instances) [Mast]. These servers run Mastodon software and communicate using ActivityPub [Activ], an open standard for federated social networking. Servers listed in Mastodon's official server directory commit to the Mastodon Server Covenant, which includes daily backups, having an emergency administrator, and giving users at least three months' notice before shutdown. Minds is mainly a centralised platform but supports ActivityPub interoperability with other networks.

SimpleX Chat is a reference application built on the broader SimpleX platform. The platform uses a decentralized client-server architecture: message transport is provided by a federation-like set of independent SimpleX servers operated by different entities, with no mandatory central authority. Endpoints communicate with servers, not directly with other endpoints for routine delivery. This architecture combines distributed control and operator diversity with server-mediated asynchronous messaging, reliability, and queue-based routing [SimpleX].

2.4 Privacy and security

Registration on most platforms requires either an email address or a mobile number. This can reduce anonymity, as identities may be revealed through other activities linked to these credentials. Another risk to anonymity is the logging of user IP addresses. This can be mitigated by using a virtual private network (VPN). A VPN establishes an encrypted connection between the user's device and a remote server operated by the VPN provider and masks the user's IP address from the destination server. However, the VPN provider can still observe user traffic.

4chan allows the optional use of tripcodes to create persistent pseudonyms without registration. A tripcode is a hashed version of a password that allows users to be recognised across posts. A user enters a name followed by a password. The platform applies a hash function to the password and generates a string of characters called a tripcode. The same input produces the same output. The name and tripcode are displayed on posts, enabling other users to recognise contributions from the same individual without revealing the password.

Another risk to anonymity is tracking cookies. These cookies collect information about browsing behaviour, such as websites visited and interactions performed. When combined with other data, they may contribute to user profiling and identification. Some platforms, such as Endchan, state that they do not use tracking cookies.

Some platforms offer account verification services. Verified accounts are presented as more authentic, may receive greater visibility, and can reduce impersonation risks. Verification does not necessarily improve anonymity.

Some platforms, including Endchan, allow posting via anonymising networks. Anonymisers conceal users' IP addresses and network metadata by routing traffic through intermediary nodes. Commercial services such as VPN providers offer this function, while open-source networks such as Tor, I2P, and LLARP provide decentralised anonymity routing. These tools protect network-level metadata. Encryption of message content depends on the application layer. Tor is required to access Tor onion services on the dark web, which are not indexed by conventional search engines [Berg23].

With end-to-end encryption (E2EE), a message is encrypted on the sender's device and decrypted only on the receiver's device. Only the communicating parties can read the content. Hypertext Transfer Protocol Secure (HTTPS) uses the Transport Layer Security (TLS) protocol to encrypt data during transmission between a user's browser and a web server. HTTPS protects data in transit but does not provide end-to-end encryption since the server operator can access the content.

All Signal messages and calls are end-to-end encrypted by default using the Signal Protocol, which combines a key agreement protocol (including PQXDH) with the Double Ratchet algorithm. This prevents Signal from accessing message content, and the service is designed to collect and retain minimal metadata [Sign1], [Sign2].

Regular Telegram chats are not end-to-end encrypted. Telegram uses client-server encryption for standard chats, which protects data in transit between the user's device and Telegram's servers. Telegram can access the content of these messages. End-to-end encryption is available only in Secret Chats, which also support additional privacy features such as self-destruct timers and restrictions on message forwarding.

Odysee is built on the LBRY protocol. LBRY stores content metadata and claims on a blockchain, while the video files are distributed through a P2P network. Blockchain records are difficult to alter, but the Odysee website and apps are centrally operated interfaces that can delist or restrict access to content.

SimpleX does not require a phone number or a global user identifier at the protocol level. User profiles are local to client devices, and contact establishment occurs out of band, for example through QR codes or

invitation links. This reduces cross-contact metadata linkability and avoids a central identity namespace. Optional two-hop private routing relays messages through an intermediary SimpleX Messaging Protocol (SMP) server before reaching the recipient's server, which helps conceal sender IP information from the destination server. TLS protects the transport layer, while SMP and agent-layer cryptography provide end-to-end confidentiality and integrity. Fixed-size message framing reduces traffic pattern leakage, and Tor operation is supported for stronger metadata protection [SimpleX].

Some far-right platforms, including The Daily Stormer, Gab, and BitChute, accept donations in cryptocurrencies such as Bitcoin. All on-chain Bitcoin transactions are publicly recorded on the blockchain and can be analysed by anyone with access to the ledger [Naka06]. This transparency allows funds to be traced between pseudonymous wallet addresses. Although wallet addresses are not directly linked to real-world identities, they can be associated with individuals through additional information or blockchain analysis. Off-chain transactions, such as those conducted on the Lightning Network, occur within payment channels and are not recorded individually on the main blockchain; only the transactions that open and close a channel appear on-chain as final settlement [PoDr16]. Because Bitcoin transactions remain publicly traceable, The Daily Stormer later shifted to requesting donations in Monero, a cryptocurrency designed to conceal sender, recipient, and transaction details [Kine21].

2.5 Business model

Social networking platforms require revenue to cover operational costs and to sustain content production. Many platforms offer incentives to encourage creators to contribute and retain audiences.

Platforms generate income through advertising, subscriptions (including premium accounts), in-app purchases or paid add-ons, donations, crowdfunding, affiliate partnerships, data licensing or API access, transaction fees on creator earnings, and sponsorships.

Reddit premium subscription offers ad-free browsing, exclusive avatar gear, custom app icons, and access to r/lounge [Redd1].

In addition to advertising, X generates revenue through data licensing and subscription services such as X Blue. X licenses access to public data and APIs to companies and developers, who use this data for analytics, research, and integration into third-party tools. X Blue is a subscription service that provides access to additional features and benefits.

Mastodon does not include built-in monetisation mechanisms in its software. Individual server operators determine how to fund their services. Some offer paid accounts, while others rely on donations or crowdfunding platforms such as Patreon (https://www.patreon.com/).

Gab generates revenue through subscriptions, donations, and affiliate partnerships. In affiliate marketing, the platform earns a commission when users purchase products or services through referral links shared on the platform. The commission is typically a percentage of the sale.

Creators on some platforms can earn income directly from audiences or through platform-based monetisation programmes. For example, YouTube allows users to earn money if they are accepted into the YouTube Partner Programme [Yout1]. Revenue sources include advertising, YouTube Premium revenue sharing, channel memberships, integrated shopping features, Super Chat and Super Stickers, and Super Thanks. Super Chat and Super Stickers allow viewers to pay to highlight messages or animated images during live chats. Super Thanks allows viewers to pay for an animation and a highlighted comment under a video or Short.

2.6 Moderation

Social networking platforms use moderation and filtering to remove harmful and illegal content. Unauthorised content can include material that incites violence, terrorism, racism, antisemitism, homophobia, and transphobia. All platforms have terms of service that prohibit such content to various degrees. Users who breach these terms can have their accounts suspended or terminated. There are three types of moderation:

Human moderation: most platforms use in-house or outsourced employees to review and remove unauthorised content. Some platforms empower super-users to directly remove unauthorised content. Others require users to report such content, which is then reviewed by platform moderators before a removal decision is made. Finally, external entities such as the EU Internet Referral Unit [3]or trusted flaggers [4]can request removal of Internet content. The EU Internet Referral Unit targets violent extremist online content materials. Trusted flaggers are special entities under the Digital Services Act (DSA) [DSA22] who are experts at detecting hate speech and terrorist content.

Automated moderation: an increasing number of platforms are using automated tools to detect and remove unauthorised content. One immediate benefit of automated tools is that they are much faster than humans. Moreover, unlike human moderators, they are proactive and can block unauthorised content even before it is published. This capability is becoming critical for avoiding liability under new European laws that require rapid action [Elki20]. For example, the DSA mandates that "In order to benefit from the exemption from liability for hosting services, the provider should, upon obtaining actual knowledge or awareness of illegal activities or illegal content, *act expeditiously* to remove or to disable access to that content". Another advantage is that they can shield human moderators from the psychological harm associated with viewing violent or abusive content [Vall23]. Platforms that use automated tools to detect unauthorised content include Discord, Facebook, Reddit, Truth Social, X/Twitter, and YouTube.

Automated moderation uses metadata filtering, hash algorithms, and machine learning algorithms.

Most digital files contain information about their content, known as metadata. Metadata filtering tools use metadata to identify content that meets specific criteria. However, the effectiveness of these tools is limited because a file's metadata can be easily manipulated.

A hash algorithm assigns a fixed-size binary string (hash) to an input image or video. The hash is significantly smaller than the original file. The hash algorithm should ensure that the probability that two different files have the same hash is very low. Hashes of known prohibited content are stored in databases. Whenever a new file is uploaded to the platform, its hash is calculated and compared against the database hashes. Since hashes are much smaller than the original files, maintaining and searching the database is more efficient that with the original files. One hash-based automated tool to identify extremist content is eGLYPH [Coun18].

Machine learning algorithms, such as DNNs, can be used to detect harmful content within images. Machine learning algorithms are trained on large datasets that are manually labelled by human annotators. However, the accuracy of these labels may vary, as identifying objectionable content is inherently subjective. The accuracy of these DNNs largely depends on the quality of the training datasets. Moreover, many DNNs have limited generalizability to new, unseen data. Natural language processing (NLP) models are used to detect hate speech, extremist content, and conduct sentiment analysis [Jaha23]. These NLP models are also trained on datasets that have been annotated by humans, which presents challenges in ensuring annotation quality

[3] https://www.europol.europa.eu/about-europol/european-counter-terrorism-centre-ectc/eu-internet-referal-unit-eu-iru. Accessed: 29/08/24.
[4] https://digital-strategy.ec.europa.eu/en/policies/trusted-flaggers-under-dsa. Accessed: 29/08/24.

and consistency [Sang22]. When a model is provided with a collection of documents, known as a corpus, it identifies patterns and features associated with each annotated category.

Singh [Sing19] provides a comprehensive overview of content moderation practices on Facebook, Reddit, and Tumblr.

Hybrid moderation: both human and automated tools are used. For example, automated tools are used to flag suspicious content, which is then reviewed by human moderators.

Some platforms, including Facebook, allow users to block certain words and ban people.

3 Deepfakes

Deepfakes are audio, images, and videos altered using DNNs. Deepfakes have been used to defame, disinform, and incite to violence, posing threats to democracy and social stability [Pawe22], [Busc23]. Examples include President Zelinsky calling on Ukrainians to surrender [Wake22], President Biden's announcement of the return of conscription to address tensions with Russia and China [Thal23], Ocasio-Cortez's video call with Pelosi and Biden [Reut23], and Russian state TV broadcast of Ukraine's security chief admitting Kyiv's role in the terrorist massacre at a conservatory in Moscow [Clea24]. While the quality of these deepfakes was poor, limiting their impact, rapid advances in deepfake technology and growing accessibility could soon change the situation.

The fast progress of deepfake technology has led to a growing interest in the field. In recent years, several comprehensive surveys on deepfake generation and detection have been published [Tolo20], [Mirs20], [Alma21], [YXFL21], [Altu22], [Nguy22], [Pass24], [Rana22], [Seow22], [Zhan22], [Abba24]. This report will not try to replicate these works. Instead, the report aims at providing a more accessible overview of key developments. Moreover, the report will explore the latest advancements that have emerged since the publication of previous surveys. The goal is to offer a resource for policymakers, practitioners, and the general public who want to stay informed about this rapidly evolving field.

In deepfakes, DNNs can be trained for entire face synthesis, facial reenactment (transferring facial expressions or body motion from one person to another), face manipulation (e.g., changing age, gender, hair style), and face swapping (replacing a face by another face).

The main DNNs used to generate deepfakes are autoregressive models, autoencoders, generative adversarial networks (GANs), and diffusion models. Autoregressive models generate pixels sequentially by predicting each pixel value based on the values of previously generated pixels. The network is trained to learn to make these predictions. Typical DNNs used include recurrent neural networks (RNNs). Autoencoders consist of an encoder network and a decoder network. The encoder generates a latent hidden representation from the input data, while the decoder uses this representation to reconstruct the original data. The latent hidden representation captures the most important features and underlying structure of the data while reducing its dimensionality. Among the most popular autoencoders are variational autoencoders (VAE). VAEs make strong assumptions about the probability distribution of the latent variables. A GAN consists of a generator network and a discriminator network. The generator synthesises videos, and the discriminator evaluates them. The generator is trained to fool the discriminator, while the discriminator is trained to determine whether the output of the generator is a real video or a deepfake. The two networks are iteratively optimised until they reach a state where the discriminator is unable to distinguish the synthesised video from a real video. Diffusion models consist of a forward process and a backward process. The forward process gradually converts images into a simple probability distribution, such as Gaussian noise. The backward process uses a trained DNN to reverse the forward process. To generate an image, the model starts with random noise and iteratively refines it into a coherent image using the backward process. Diffusion models typically use convolutional neural networks

(CNNs) such as U-Nets and residual networks (ResNets) in conjunction with transformer-based networks such as vision transformers (ViTs).

Many deepfake software applications are publicly available [Altu22], [Seow22]. Some of these applications can be used with very little technical skill. Popular applications include FaceApp (faceapp.com), Stable Diffusion (https://stability.ai/stable-image), DALL.E (the latest version being DALL.E 3: DALL·E 3 | OpenAI), StyleGAN2 (https://github.com/NVlabs/stylegan2), DeepFaceLab (https://github.com/iperov/DeepFaceLab), and ZAO (https://zaodownload.com/). FaceApp allows facial attribute manipulation. DALL-E is a diffusion-based model developed by OpenAI that generates images from text descriptions. Stable Diffusion is another text-to-image network. It is based on a VAE and a U-Net. StyleGAN2 synthesises images using GANs [Karr20]. DeepFaceLab is an open-source framework for video face swapping. ZAO is a popular free Chinese application that allows users to superimpose faces onto TV and movie characters. The user only needs to provide several photos displaying different facial expressions.

Training DNNs is time-consuming and requires large datasets. Existing datasets include DeepFakeTIMIT (620 videos with faces swapped), UADFV (49 real videos and 49 deepfake videos), FaceForensics++ (1,000 original face videos, and 4,000 deepfakes obtained with two computer graphics methods and two deep learning methods), Celeb-DF (5,639 deepfake videos), DFDC (23,654 real face videos of 960 subjects, and 104,500 deepfakes obtained with eight different methods), and DeeperForensic (50,000 real videos and 10,000 manipulated videos) [Zhan22],[Altu22].

Deepfake quality can be evaluated through user studies involving human participants using statistical metrics such as accuracy (Acc) and area under curve (AUC). Given a set of real images and deepfakes, one asks each participant whether the data is real (class negative) or fake (class positive). Then, one counts the number of true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN). TP is a fake image classified as fake. FN is a fake image classified as real. FP is a real image classified as fake. TN is a real image classified as real. Accuracy is the ratio (TP+TN)/(TP+FN+FP+TN). When the user gives a confidence score that the image is fake, a TP occurs when a fake image is given a confidence score greater than or equal to a given threshold. An FN occurs when a fake image is given a score smaller than the threshold. An FP occurs when a real image is given a score greater than or equal to the threshold. A TN occurs when a real image is given a score smaller than the threshold. AUC is the area under the receiver operating characteristic (ROC) curve. The ROC curve gives the true positive rate (TPR) as a function of the false positive rate (FPR) when the threshold is varied. Here, TPR= TP/(TP+FN) and FPR= FP/(FP+TN). AUC is between 0 and 1. The closer it is to 1 the better the classification.

In [Poco23], 260 participants were shown 20 images (10 real images from the Internet and 10 deepfakes generated by Stable Diffusion or DALL-E 2) and asked whether the images were real or fake. The accuracy was only 61%. In another user study [Nigh22] involving 315 participants and image deepfakes generated with StyleGAN2 (https://github.com/NVlabs/stylegan2), the accuracy was even lower (48.2%). Figure 7 shows images of the most and least accurately classified images from the study. These studies indicate that image deepfakes are now undistinguishable from real images.

Video deepfakes have not reached that quality level. When 100 participants were asked to assess how real videos from seven deepfake datasets are on a scale from 1 to 5 (1 – clearly disagree, 2 – weakly disagree, 3 – borderline, 4 – weakly agree, 5 – clearly agree), the percentage of ratings greater than or equal to 4 was only 8.4%, 12.3%, 14.1%, 21.9%, 23%, 61%, and 64.1% [Altu22].

As human studies are time-consuming and expensive, several automated deepfake detection methods have been proposed. The most successful ones are based on DNNs. A study in [Maia24] compared the performance of two DNNs to human performance (120 online participants) on 24 deepfake photos and 24 real photos. The

participants were asked to rate their confidence on a scale from zero to six. ResFormer outperformed human detection, achieving a higher accuracy rate of 93.7% compared to 71.61%, and a larger AUC of 0.938 compared to 0.707. Also, no participant outperformed ResFormer. However, there were three deepfake photos that were misclassified by ResFormer but correctly classified by 54.17%, 56.67%, and 72.5% of the participants, respectively. Note that human accuracy was higher than in [Poco23] and [Nigh22], showing the dependence of the results on the dataset. The study also showed that the performance of AI systems is superior to human performance when a model is trained and tested on images generated with the same family of techniques.

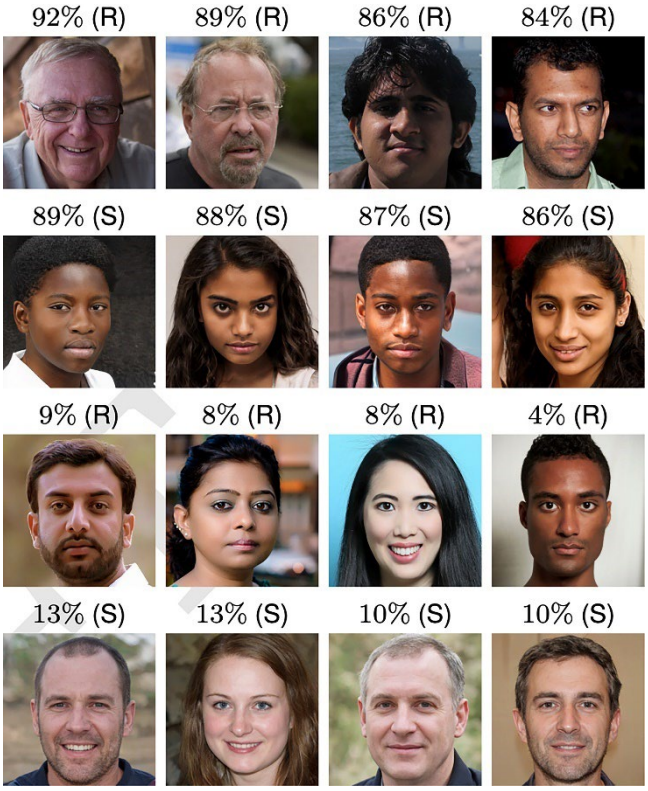

Fig. 7: Most and least accurately classified real (R) and synthetic (S) images from the user study [Nigh22].

Mahara and Rishe [MaRi26] categorizes deepfake image detection methods into seven main groups. The first group is spatial-domain analysis, which looks at pixel-level content and searches for artifacts such as unnatural textures, edge irregularities, and colour inconsistencies introduced during synthesis. The second group is frequency-domain analysis, which transforms images into the spectral domain to detect abnormal periodic patterns, noise behaviour, and frequency components that often appear in generated images.

The third group is fingerprint analysis. These methods try to detect traceable signatures left by the image generation process.

The fourth group is patch-based analysis, which examines local image regions instead of the whole image. This helps detect subtle artifacts that may be visible only in small areas.

The fifth group is training-free detection. These methods do not depend on supervised training with real and fake datasets. Instead, they rely on statistical inconsistencies, analytical cues, or reconstruction behaviour in the image representation.

The sixth group is multimodal and reasoning-based detection. In this category, VLM-based approaches use image-text alignment to identify inconsistencies through cross-modal embeddings, usually without explicit

reasoning chains. MLLM-based approaches go further by combining visual and textual inputs to produce both a detection decision and an explanation of that decision.

The seventh group is commercial detection frameworks, including proprietary approaches such as invisible watermarking signals embedded during generation and later used for verification and detection.

To compare methods, the paper evaluates performance on public datasets and checks generalizability across three criteria: cross-family, cross-category, and cross-scene. Cross-family asks whether a detector trained on one generator family can still detect images from a different generator family. Cross-category tests transfer across semantic classes, such as training on faces and testing on animals. Cross-scene tests robustness across datasets with different data distributions.

The main pattern in the results is that multimodal frameworks tend to be more robust and adaptable across generator shifts. Spatial and frequency methods remain important because they capture core low-level forensic cues, but they can lose strength under compression and higher visual realism. Fingerprint and patch-based methods improve sensitivity and cross-generator robustness by using pipeline traces and local texture cues. Training-free methods improve scalability because they avoid retraining, but challenges remain in interpretability and resistance to adversarial manipulation. Reasoning-based multimodal models improve explainability and semantic grounding, but they come with high computational cost.

Based on these results, the paper suggests that hybrid frameworks are a strong direction for future work. The idea is to combine the semantic reasoning power of multimodal models with the efficiency of training-free approaches to build detection systems that are robust, interpretable, and usable in real time. Performance is mainly reported using classification accuracy and average precision (AP).

Xie et al. [Xie26] frame DeepFake detection as a binary classification task: decide whether a media sample is real or fake. The paper groups existing methods into cue-based methods and data-driven methods. Cue-based methods rely on forensic traces left by imperfect generation. Data-driven methods learn discriminative representations from training data.

In cue-based detection, the core assumption is that even strong generators still leave artifacts. These methods look for inconsistencies in biological signals and in spatial, temporal, and frequency domains. Biological-signal methods use cues such as blink patterns, eye characteristics, pulse-related signals, and behaviour. These cues are hard to reproduce consistently, but extraction can fail under poor lighting, occlusion, motion jitter, or low-quality capture. Spatial methods inspect image-level and pixel-level artifacts, including noise patterns, forensic fingerprints, quality distortions, and lighting inconsistencies. Their weakness is that newer generators suppress obvious spatial artifacts, which reduces separability. They also tend to generalise poorly to unseen generators and are more reliable for image-level detection than full video-level decisions. Temporal methods analyse frame-to-frame continuity, motion consistency, and facial dynamics. They can capture temporal anomalies, but they are computationally expensive and often sensitive to video quality. Their performance can also drop when manipulations are subtle or when artifacts differ from those seen during training. Frequency methods detect spectral anomalies, such as unrealistic high-frequency textures or abnormal low-frequency smoothness. These methods are useful, but they also struggle with subtle edits and cross-model generalization.

The paper divides data-driven detection into four subtypes. Generic network methods use standard deep models, with or without attention, to learn fake-vs-real representations. Data-augmentation methods improve diversity by adding synthetic or transformed samples during training. Multi-modal methods learn joint representations across image, audio, and video to expose cross-modal inconsistencies. Multi-supervision methods use richer supervision signals to improve transfer to complex scenarios.

The survey reports a large benchmark over 36 detectors from both families, evaluated in within-domain and cross-domain settings. Within-domain means training and testing on the same dataset. Cross-domain means training and testing on different datasets. AUC is the main metric throughout. In within-domain tests on first-generation datasets, many methods perform extremely well, with advanced models above 0.99 AUC and some reaching 1.00. On second-generation datasets, several methods still exceed 0.90 AUC. On third-generation datasets, performance drops substantially, with reported AUC values around 0.6960, 0.7330, 0.7230, and 0.7130. The paper links this drop to harder forgeries from newer models and broader generation settings.

Cross-domain results are consistently weaker. Performance declines for all methods, especially across dataset generations. The paper attributes this to distribution shift between older and newer generator families and to heavy dependence on supervised training data. Detectors tuned to specific artifacts in seen datasets do not transfer well to unseen generators. The authors highlight that this is especially visible with newer model families, including Midjourney, DALL·E, and Stable Diffusion.

The main conclusion from the paper is that generalization remains the primary bottleneck. Current detectors can score very high in favourable in-domain settings but lose reliability on unseen data and newer synthesis pipelines. The paper also emphasises two major robustness threats: adversarial attacks and media degradation from compression and noise.

The paper [Li25] benchmarks 17 detectors and 10 vision-language models (VLMs) on the Real-World Robustness Dataset (RRDataset). It also includes a large human study with 192 participants to test few-shot learning in AI-generated image detection. In few-shot learning, models are shown a few labelled examples of real vs. AI-generated images, then tested on new images. RRDataset contains high-quality images from seven major scenarios: war and conflict, disasters and accidents, political and social events, medical and public health, culture and religion, labour and production, and everyday life. The benchmark evaluates robustness in standard detection, internet transmission, and re-digitization settings. Internet transmission robustness measures detector performance after multiple rounds of sharing on social media platforms. Re-digitization robustness measures performance after four different re-digitization methods. RRBench is a comprehensive benchmark built on these settings and includes the 17 detection methods and 10 VLMs. None of the 17 detectors reaches saturated performance on RRDataset, and the best accuracy is only 89.59%. Network transmission and re-digitization significantly reduce performance, showing weaknesses that are often missed in conventional evaluations.

The most robust method uses a hybrid design that combines multiple expert models and feature types. The highest overall performance on RRBench is achieved by a method that combines diffusion-based image reconstruction with contrastive training.

The study also presents a large human benchmark, with 192 participants and 240 test images. Human accuracy drops sharply on transmitted and re-digitized images. However, after few-shot learning, this transmission and re-digitization effect is effectively reduced.

The results show clear limits in current detection methods under real-world conditions and highlight the value of human adaptability for developing more robust detection approaches.

4 Bots

Bots can be used by extremists for propaganda generation, providing know-how, platform boosting, and adversarial platform flooding [Bael24], [Hick23].

Generative chatbots can be used to automatically generate extremist content, which may contribute to radicalisation of human users. For example, the 4chan/pol forum is used for sharing racist texts generated by

fine-tuned racist versions of Meta's LlaMA model [Bael24], as well as antisemitic images and racist memes generated by Bing Image Creator [LeKo23]. For example, one of the generated memes shows male immigrants of colour with knives chasing a white woman [LeKo23]. Generative chatbots can also be exploited by extremists to provide expertise that would otherwise be difficult to obtain. For example, Lakomy [Lako23] demonstrated how ChatGPT and BingChat can be used to provide instructions and computer code for evading online detection. The same study also revealed that these chatbots can be exploited to learn about violent tactics and weaponry. Meta's LLaMA and LlaMA 2 are open-source language models and can be modified to remove built-in safeguards. AI-risk company Palisade Research created a rogue version of the LlaMA 2, which gave advice on cybercrime, building weapons, writing hate speech, and planning homicide [Bael24].

Platform boosting involves using AI bots to artificially inflate the popularity of individuals on social media platforms. Forum spambots search the web, looking for guestbooks, wikis, blogs, forums, and other types of web forms where they can submit fake content. They may use automation and CAPTCHA-solving methods to bypass protections. The aim is to create a false sense of importance, which can attract real users and enhance the reputation of extremist leaders within online communities. Other spam messages are not meant to be read by humans but are instead posted to increase the number of links to a particular website, boosting its search engine ranking.

Fake views generated by bots can inflate view counts and make videos appear more popular than they are. If not detected, this can distort engagement signals used for ranking and recommendations, which may increase the reach of misleading content. YouTube monitors for invalid traffic and may adjust view counts by removing views it identifies as artificial or automated [Cast24].

Adversarial platform flooding involves using synthetic content to overwhelm an enemy or rival platform. Adversarial platform flooding can decrease the attractiveness of the platform, delegitimise the platform, shift users' views, and even make the platform unusable [Bael24].

5 Emerging Technologies

This section examines emerging, still-maturing technologies that could pose a near-future threat if adopted by far-right actors for propaganda and mobilisation. It focuses on five areas: audio-visual deepfakes, synthetic identities, hyper-personalised targeting, affective computing, and VR/AR immersive influence.

Recent literature shows a clear shift from basic lip-sync deepfakes to broader audio-visual identity cloning. New systems not only match mouth motion to speech, but also preserve identity, control emotion, and maintain stable video quality over longer clips. Diffusion-based and 3D-aware methods are among the strongest current approaches in this area. Systems such as EmotiveTalk, Teller, InsTaG, and AudCast report major improvements in expressiveness, temporal stability, and realism in audio-driven portrait or human video generation [Wang25], [Zhen25], [LiZh25], [Guan25]. These systems indicate that one image plus driven audio can, in many settings, produce convincing talking videos with consistent facial dynamics and natural speech timing.

EmotiveTalk [Wang25] focuses on controllable emotional speaking style and more stable long-duration output, which addresses a common weakness of earlier models. Teller [Zhen25] focuses on real-time streaming generation, which reduces latency and makes live or near-live synthetic video more feasible. InsTaG [LiZh25] shows that short source clips can be enough to personalise a 3D talking head, which lowers the data barrier for realistic impersonation. AudCast [Guan25] extends this to full human video generation with cascaded diffusion transformers, improving motion richness and visual coherence.

Recent research also shows fast progress in voice cloning quality, including speaker similarity, prosody, and cross-lingual transfer. This supports multilingual impersonation where the same cloned identity can speak in

different languages while retaining recognizable voice traits. When paired with modern talking-head models, this creates stronger end-to-end impersonation: cloned voice, synchronized face, and controlled emotion [AzEl25].

Beyond synthetic media generation, current work highlights a second shift: persistent synthetic personas that can influence over time. Synthetic identities and AI influencers are a distinct class of information threat. A synthetic identity is not just one fake post or one fake image; it is a persistent persona that combines profile history, language style, network behaviour, and often synthetic voice or video output. Recent literature shows that these personas can appear coherent over long periods and across platforms. This continuity can increase credibility and can make detection harder than detection of isolated deepfake items [WiPr24], [Mose25].

Current systems build these personas from several AI components. Large language models generate consistent text in a stable personal style. Image generators produce profile photos and lifestyle content. Voice cloning and talking-head generation add audio-video presence. Scheduling and automation tools maintain regular posting patterns and interaction cycles. The result is a believable account that can answer, adapt, and persist. This differs from older bot models, which often showed repetitive language and weaker personalization [Stoc24], [RoMaGo25].

Research on virtual influencers and parasocial interaction shows that users can form social attachment to media personas, including artificial ones. Repeated exposure, emotional tone, and perceived familiarity can increase message acceptance. This effect does not require users to believe the persona is human in every case. It requires a stable identity that feels socially present and relatable. In hostile settings, this mechanism can support recruitment and ideological grooming through gradual relationship building [StLiAn22], [LiWa25], [BrBe25].

Recruitment and coordination risks increase when synthetic personas operate in groups. Many aligned personas can amplify each other, simulate majority opinion, and pressure real users through repetition and social proof. This can shift perceived norms and create false impressions of broad support. In extremist or conspiratorial environments, these dynamics can lower barriers to entry for new recruits and strengthen in-group identity [Schr26], [Stoc24].

Attribution remains a main challenge. A single synthetic persona can use different channels, languages, and media formats while keeping one narrative identity. Traditional moderation often focuses on content-level violations in single posts. That approach can miss campaign-level structure. Recent policy and technical literature recommend network-level analysis, provenance signals, and cross-platform coordination detection. These methods aim to identify patterns of orchestration, not only false statements in individual messages [Stoc24], [Mose25], [RoMaGo25].

Debunking is also difficult. Fact-checking may correct one false claim, but synthetic personas can quickly post new versions, reframe the correction, or move to a different topic. They can also combine accurate facts with false claims, which makes their content seem credible. As a result, one-time corrections often have limited effect. Proposed responses include stronger identity verification on platforms, clear labels for synthetic media, targeted limits on rapid mass sharing of high-risk content, and early detection of coordinated persona networks [WiPr24], [Mose25].

Taken together, synthetic identities and AI influencers can build trust, recruit, and coordinate more effectively than earlier bot networks because they sustain believable social presence over time and across modalities. They are harder to attribute because control can be distributed and anonymized. They are harder to debunk because manipulation occurs at network and relationship level, not only at claim level.

A third shift concerns delivery optimization. Hyper-personalised algorithmic targeting uses data about individuals to adapt messages to personality, preferences, context, and likely vulnerabilities. This approach can increase persuasion and engagement compared with generic messaging. Effects are often stronger when the message matches psychological traits and situational cues [SiEdLe24].

Matz et al. [Matz24] show that LLM-generated personalised messages can outperform non-personalised messages in persuasion tasks. The study also shows scalable message generation. This lowers the cost of producing many message variants for many audience segments and moves personalization from manual campaign design to automated optimization.

Simchon et al. [SiEdLe24] find that personality-congruent political ads are more persuasive than non-targeted ads in several experiments. Their work also shows that AI methods can automate this process across large audiences. This strengthens concerns about microtargeting in elections and public debate because message influence can be tuned to individual psychology rather than public argument. Related evidence suggests that simple transparency warnings may not remove this advantage. Carrella et al. [Carre25] report that warnings about personality-based targeting do not reliably remove the persuasive effect of targeted political ads.

Recent platform-focused studies add behavioural evidence on engagement. Dekker, Baumgartner, and Sumter [DeBaSu25] examine TikTok feed personalization and show that personalization changes user behaviour and experience in measurable ways. This supports the broader claim that recommender systems can shape attention allocation and interaction intensity, not only content order. Consumer research also reports a dual effect: personalization can improve relevance and short-term response, but it can increase privacy concern and fatigue when users feel over-targeted [HuLi25], [HaMa24].

Targeting does not affect all users equally. Effects depend on digital literacy, persuasion knowledge, emotional state, and context. Users with lower awareness of algorithmic influence may show weaker resistance to personalised persuasion. Moravec et al. [Mora25] argue that knowledge gaps in algorithmic personalization reduce user ability to identify manipulation and manage exposure.

The field does not claim universal or unlimited effects. Some studies report heterogeneous or modest average impacts in specific setups. Hackenburg and Margetts [HaMa24] find persuasive effects for AI-generated political messages but limited aggregate gains for microtargeting in their design. This suggests that effect size depends on platform context, outcome measure, targeting quality, and audience state.

A fourth shift concerns emotional inference. Advances in affective computing and emotion recognition now support finer inference of user state from text, voice, face, and physiology. Recent work shows a shift from unimodal systems to multimodal systems that fuse signals with transformers, attention mechanisms, and generative models [Geet24], [Ma25], [Alsa24]. This shift improves performance on benchmark datasets, but it also increases the risk of overclaiming real-world accuracy. Controlled datasets often do not match real deployment conditions, where lighting, language, culture, noise, and context vary sharply [Geet24], [Pan23].

The latest methods use multimodal fusion to combine facial expression, speech prosody, text semantics, and physiological data. Recent work also uses modality-aware attention and graph-based fusion to model interactions across channels [DePoPe25], [Woo25]. However, a model can perform well in one dataset and fail in another setting with different demographic mix, recording conditions, or task framing [Geet24], [Alsa24].

Generative approaches for affective computing are growing rapidly, including data synthesis, representation learning, and augmentation for sparse emotion labels [Ma25]. These tools can improve training, but they can also hide uncertainty if synthetic data quality is not audited. If training data or labels are biased, the system can produce confident but wrong emotion outputs. This risk is high when emotion inference supports high-impact decisions or targeting systems [Ma25].

Emotion recognition can support adaptive persuasion systems that tune message timing, framing, and intensity to inferred user state. If systems infer stress, anger, or fear, targeting logic can exploit those states to raise compliance or engagement. Studies on personalized algorithmic decision-making and AI deception risks warn that such systems can influence behaviour in ways users do not detect, especially when transparency is weak [Lu24], [WiPr24]. In this context, overstated accuracy is not a minor technical issue. It can justify intrusive or manipulative interventions that rest on uncertain inference.

A second risk is escalation through feedback loops. If a system predicts high arousal and then serves more emotionally intense content, it can reinforce the same state and increase polarization or distress. The literature on AI-enabled influence operations and personalization risks supports this mechanism, even when single-model accuracy is moderate [Stoc24], [Mora25]. The danger is cumulative optimization: repeated micro-adjustments can outperform one-off persuasion attempts.

A fifth shift concerns immersive interaction environments. VR/AR and immersive metaverse environments can increase influence power because they combine presence, embodiment, and real-time social interaction. Recent literature shows that immersive systems can produce stronger psychological involvement than standard social media. Users can feel being there, being with others, and being this avatar. These states can change attention, emotion, and social behaviour in ways that can increase compliance with group norms and persuasive cues [HaBa24].

A core mechanism is social presence. In social VR, voice, gaze, distance, and gesture are available in real time. This gives interactions higher immediacy than text feeds. Studies show that social norms from physical settings can transfer into virtual environments, including conformity signals and interpersonal regulation. When users perceive strong co-presence, they often respond as if the social situation is real. This can make pressure, persuasion, and status signalling more effective [HaBa24], [Han24].

Users can identify with avatars and experience changes in self-perception and behaviour. Recent work shows that perspective, avatar form, and behavioural realism affect trust, disclosure, and interaction outcomes. If an attacker controls a realistic avatar with synchronized speech and gesture, the target may assign higher credibility and social authority to that persona. This can intensify manipulation in recruitment or coordinated influence settings [SoTiRu25].

Voice tone, pauses, body movement, and proxemics provide rich signals for persuasion. In immersive environments, these cues can be optimized during interaction. This supports adaptive influence tactics: message framing can change in response to user reactions, group mood, or resistance. Communication research on social VR and immersive persuasion highlights this shift from static content persuasion to interactive behavioural shaping [HaBa24].

However, current evidence does not support simple claims that immersion always increases persuasion. Effects vary by design quality, context, and user characteristics. Stronger immersion combined with live social interaction can increase the potential for social pressure and manipulative escalation. This is most concerning when platforms combine realistic avatars, persistent identity, and real-time voice and gesture channels with weak moderation or weak provenance controls.

Recent news shows early use of synthetic identities by the far right. *The Guardian* reported in January 2026 that an AI-generated schoolgirl avatar (“Amelia”) was repurposed and amplified by far-right accounts across mainstream platforms [Quin26]. There is also evidence of hyper-personalised targeting. A February 2026 *Guardian* investigation [Down26] described AI deepfake fraud on an industrial scale, including content targeted at individuals and tailored, personalised scams. This shows that personalised message adaptation is already operational in malicious AI campaigns.

6 Conclusion

Most platforms allow pseudonym registration, obscuring user identity but not entirely preventing tracking through IP addresses, email addresses, or phone numbers.

The increasing prevalence of end-to-end encryption for messaging and video calls offers a degree of privacy, which can be further enhanced through external services like VPNs. However, most platforms operate on centralized servers, posing significant privacy risks. Only a small number of anonymity networks, such as Tor and Lokinet, can reduce traceability by routing traffic through relay networks, but they do not guarantee complete anonymity, and end-to-end encryption still depends on the application layer (e.g., HTTPS).

The business models of these platforms, primarily driven by advertising, donations, crowdfunding, and premium subscriptions, influence their approach to content management. While generating revenue, these models may also incentivise tolerance of extremist content. Moreover, the potential for privacy breaches through data collection and the emerging use of blockchains add further complexities. The ability of content creators to monetise through tips, donations, and cryptocurrencies provides additional avenues for extremist groups to fund their activities.

Deepfakes, manipulated audio, images, and videos created using DNNs, pose significant threats to democracy and social stability by enabling disinformation, defamation, and incitement to violence. Although current deepfakes often suffer from low quality, rapid technological advancements may soon increase their impact. Deepfake generation methods include autoregressive models, autoencoders, GANs, and diffusion models, which are becoming increasingly accessible through free and simple to use applications. As realism improves, misuse risks increase. At the same time, deepfake detection remains a major challenge. Current detection methods can perform well in controlled settings but often lose robustness and generalisation across datasets, generator families, compression levels, and real-world deployment conditions.

Extremists use bots to disseminate propaganda, share expertise, manipulate platforms, and overwhelm opposition. Large language models are used on alt-tech platforms to generate racist content, fuelling radicalisation. These bots can also provide instructions on cybercrime and building weapons. By artificially inflating social media metrics, bots create a false sense of popularity for extremist content. Moreover, synthetic content is used to flood rival platforms to undermine their credibility and usability. To combat this, platforms scrutinise and remove bot-generated data.

Related work in Deliverable D3.2 [SMIDGE3.2] shows that recommendation algorithms strongly shape what users see and can contribute to the spread of misinformation and inflammatory content. By ranking and repeatedly resurfacing material based on engagement and predicted relevance, these systems can amplify popularity bias, encourage clustering into like-minded communities, and create feedback loops that speed up diffusion. Audits and simulations suggest that optimising for engagement or recommendation accuracy can, in some settings, increase exposure to false or polarising content.

Emerging, still-maturing technologies may further expand these risks in the near future if adopted by far-right actors. Key areas include audio-visual deepfakes, persistent synthetic identities, hyper-personalised targeting, affective computing, and VR/AR immersive influence. These technologies increase the potential for scalable, adaptive, and harder-to-attribute influence activity. They shift the threat from isolated false content toward continuous, multi-modal persuasion that combine realism, personalisation, and social immersion. This raises urgent policy and technical needs for stronger provenance mechanisms, cross-platform coordination, stronger detection methods, and safeguards against manipulative AI-enabled influence.

This report shows that effective monitoring, regulation, and platform governance depend on infrastructure as well as platform rules. End-to-end encryption limits routine content monitoring, so policy should emphasise

transparency, consistent enforcement standards, and auditable reporting rather than broad access to private communications. Decentralised or federated services further reduce accountability because there may be no single operator able to enforce rules across the whole network. In such cases, attention often shifts to supporting services that keep platforms online and funded, including app stores, hosting providers, domain registrars, and payment services. These measures can be effective, but they raise questions about proportionality, due process, and the concentration of private power over online speech. Clear standards, appeal mechanisms, cross-platform coordination, and independent auditing are therefore essential.

This report links to Deliverable D3.4 [SMIDGE3.4] by addressing the same broader problem from a different angle. D3.4 introduces the HYPE framework to explain how conspiracy theories and extremist narratives can move into mainstream spaces through influencers, communities, memes, and participatory practices. It explains how users can move along a pathway from scepticism and playful speculation to more hardened ideological positions. This report focuses on the platforms and technologies that shape those spaces.

7 References

[Abba24] Abbas, F. and Taeihagh, A., 2024. Unmasking Deepfakes: A Systematic Review of Deepfake Detection and Generation Techniques Using Artificial Intelligence. Expert Systems with Applications, 252(B), 124-260: https://doi.org/10.1016/j.eswa.2024.124260

[Activ] https://www.w3.org/TR/activitypub/. Accessed: 27/02/26.

[Alma21] Almars, Abdulqader M., 2021. *Deepfakes Detection Techniques Using Deep Learning: A Survey.* Journal of Computer and Communications, 09 (05). pp. 20-35. ISSN 2327-5219.

[Alsa24] Al-Saadawi, H. F. T., Das, B., Das, R., 2024. A systematic review of trimodal affective computing approaches: Text, audio, and visual integration in emotion recognition and sentiment analysis. *Expert Systems with Applications, 255*(Part D), 124852. https://doi.org/10.1016/j.eswa.2024.124852.

[Altu22] Altuncu, E., Franqueira, V. and Li, S., 2022. Deepfake: Definitions, performance metrics and standards, datasets and benchmarks, and a meta-review. arXiv.org. https://arxiv.org/ abs/2208.10913

[Azad23] Azad, M. W. A, Zahan, H., Taki, S. U. and Mastorakis, S., 2023. DarkHorse: A UDP-based Framework to Improve the Latency of Tor Onion Services, in 2023 IEEE 48th Conference on Local Computer Networks (LCN), Daytona Beach, FL, USA, 2023 pp. 1-6. doi: 10.1109/LCN58197.2023.10223352

[AzEl25] Azzuni, H., El Saddik, A., 2025. Voice cloning: Comprehensive survey. arXiv preprint arXiv:2505.00579.

[Bael21] Baele, S. J., Brace, L. and Coan, T. G., 2021. Variations on a Theme? Comparing 4chan, 8kun, and Other *chans*' Far-Right "/pol" Boards. *Perspectives on Terrorism*, *15*(1), 65–80. https://www.jstor.org/stable/26984798. Accessed: 26/02/26.

[Bael22] Baele, S., 2022. Artificial Intelligence and Extremism: The Threat of Language Models for Propaganda Purposes. CREST Guide, https://crestresearch.ac.uk/resources/artificial-intelligence-and-extremism-the-threat-of-language-models/. Accessed: 26/02/26.

[Bael24] Baele, S., J. and Brace, L., 2024. AI Extremism Technologies, Tactics, Actors, VOX-Pol Network of Excellence, 2024, ISBN: 978-1-911669-68-5

[Berg23] Bergman, J. and Popov, O. B., 2023. Exploring Dark Web Crawlers: A Systematic Literature Review of Dark Web Crawlers and Their Implementation, in *IEEE Access*, vol. 11, pp. 35914-35933, doi: 10.1109/ACCESS.2023.3255165.

[BrBe25] Breves, P. L., and van Berlo, Z. M., 2025. Making and breaking parasocial relationships with human and virtual influencers: An experience sampling study. *Media Psychology*, 1-29, doi: 10.1080/15213269.2025.2558029.

[Brom22] Bromwich, J. E., 2022. Before Massacre Began, Suspect Invited Others to Review His Plan, The New York Times, 17 May 2022, https://www.nytimes.com/2022/05/17/nyregion/buffalo-shooting-discord-chat-plans.html. Accessed: 26/02/26.

[Brow21a] Browning, K. and Lorenz, T., 2021. Pro-Trump Mob Livestreamed Its Rampage, and Made Money Doing It, The New York Times, 8 January 2021, https://www.nytimes.com/2021/01/08/technology/dlive-capitol-mob.html. Accessed: 26/02/26.

[Brow21b] Brown, E., 2021. Blockchain-based Odysee keeps your social media content online, ZD Net, 8 April 2021, https://www.zdnet.com/finance/blockchain/blockchain-based-odysee-keeps-your-social-media-content-online/. Accessed: 26/02/26.

[Brya20] Bryant, L. V., 2020. The YouTube Algorithm and the Alt-Right Filter Bubble, *Open Information Science*, vol. 4, no. 1, 2020, pp. 85-90. https://doi.org/10.1515/opis-2020-0007

[Busc23] Busch, E. and Ware, J., 2023. The Weaponisation of Deepfakes. ICCT Policy Brief, Dec. 2023, DOI: 10.19165/2023.2.07.

[Carre25] Carrella, F., Simchon, A., Edwards, M. *et al.*, 2025. Warning people that they are being microtargeted fails to eliminate persuasive advantage. *Commun Psychol* **3**, 15. https://doi.org/10.1038/s44271-025-00188-8.

[Cast24] Castaldo, M., Frasca, P., Venturini, T. and Gargiulo, F., 2024. Fake views removal and popularity on YouTube. Sci Rep 14, 15443. https://doi.org/10.1038/s41598-024-63649-w

[Clea24] Cleave, I., 2024. INFO WARS Russian state TV broadcasts chilling DEEPFAKE video of Ukraine spy chief 'bragging about Moscow terror attack', The Sun, Published: 16:13, 23 Mar 2024. Updated: 15:30, 24 Mar 2024, https://www.thesun.co.uk/news/26879332/russian-tv-deepfake-ukraine-moscow-attack/. Accessed: 26/02/26.

[Conr20] Conrad, P., 2020. El Paso Shooting Suspect Faces New Murder, Hate Crime Charges, Courthouse News Service, 25 June 2020, https://www.courthousenews.com/el-paso-shooting-suspect-faces-new-murder-hate-crime-charges/. Accessed: 26/02/26.

[Coun18] Counter Extremism Project, The Eglyph Web Crawler: ISIS Content on YouTube, 2018, https://voxpol.eu/file/the-eglyph-web-crawler-isis-content-on-youtube/. Accessed: 26/02/26.

[Covi16] Covington, C., Adams, J. and Sargin, E., 2016. Deep neural networks for youtube recommendations. RecSys '16: Proceedings of the 10th ACM Conference on Recommender Systems 2016; pp. 191-198. DOI: https://doi.org/10.1145/2959100.2959190

[Craw19] Crawford, B., 2022. Tracing Extremist Platform Migration on the Darkweb: Lessons for Deplatforming, Jan. 2022, https://gnet-research.org/2022/01/18/tracing-extremist-platform-migration-on-the-darkweb-lessons-for-deplatforming/. Accessed: 26/02/26.

[DeBaSu25] Dekker, C.A., Baumgartner, S.E., and Sumter, S.R., 2025. For you vs. for everyone: The effectiveness of algorithmic personalization in driving social media engagement. *Telematics and Informatics*, 101, 102300.

[DePoPe25] Devarajan, K., Ponnan, S., and Perumal, S., 2025. Enhancing emotion recognition through multi-modal data fusion and graph neural networks. *Intelligence-Based Medicine*, 12, 100291. https://doi.org/10.1016/j.ibmed.2025.100291.

[Down26] Down, A., 2026. Deepfake fraud taking place on an industrial scale, study finds. The Guardian. https://www.theguardian.com/technology/2026/feb/06/deepfake-taking-place-on-an-industrial-scale-study-finds. Accessed: 26/02/26.

[DSA22] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng. Accessed: 26/02/26.

[Elki20] Elkin-Koren, N., 2020. Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence. *Big Data & Society*, *7*(2). https://doi.org/10.1177/2053951720932296

[Endchanfaq] https://endchan.net/.static/faq.html. Accessed on 27/02/26.

[Evan19] Evans, R., 2019. The El Paso Shooting and the Gamification of Terror. *Bellingcat*. https://www.bellingcat.com/news/americas/2019/08/04/the-el-paso-shooting-and-the-gamification-of-terror/. Accessed: 26/02/26.

[Geet24] Geetha, A.V., Mala, T., Priyanka, D., and Uma, E., 2024. Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. *Information Fusion*, vol. 105, 102218. https://doi.org/10.1016/j.inffus.2023.102218.

[Guan25] Guan, J., Wang, K., Xu, Z., Yang, Q., Sun, Y., He, S., Liang, B., Cao, Y., Li, Y., Feng, H., Ding, E., Wang, J., Zhao, Y., Zhou, H., Liu, Z., 2025. AudCast: Audio-driven human video generation by cascaded diffusion transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 10678-10689.

[HaBa24] Han, E. and Bailenson, J.N., 2024. Social interaction in VR. *Oxford Research Encyclopedia of Communication*. https://doi.org/10.1093/acrefore/9780190228613.013.1489.

[HaMa24] Hackenburg, K., and Margetts, H., 2024. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, *121*(24), e2403116121, doi: https://doi.org/10.1073/pnas.2403116121.

[Han24] Han, E., DeVeaux, C., Miller, M.R., Harari, G.M., Hancock, J.T., Ram, N., Bailenson, J.N., 2024. Alone together, together alone: The effects of social context on nonverbal behavior in virtual reality. *PRESENCE: Virtual and Augmented Reality* 2024; 33 425–451. doi: https://doi.org/10.1162/pres_a_00432.

[Hern21] Hern, A., 2021. Parler goes offline after Amazon drops it due to 'violent content'. The Guardian. https://www.theguardian.com/us-news/2021/jan/11/parler-goes-offline-after-amazon-drops-it-due-to-violent-content. Accessed: 26/02/26.

[Hick23] Hickey, D., Schmitz, M., Fessler, D., Smaldino, P.E., Muric, G. and Burghardt, K., 2023. Auditing Elon Musk's impact on hate speech and bots. In *Proceedings of the international AAAI conference on web and social media* (Vol. 17, pp. 1133-1137).

[HuLi25] Huang, Y., and Liu, L., 2025. The impact of algorithm awareness on the acceptance of personalized social media content recommendation based on the technology acceptance model. *Acta Psychologica*, *259*, 105383, doi: https://doi.org/10.1016/j.actpsy.2025.105383.

[Hume19] Hume, T., 2019. Man Who Tried to Shoot Up a Mosque In Norway Was Inspired By Christchurch and El Paso, Vice News, 12 August 2019, https://www.vice.com/en/article/7x5pjz/man-who-tried-to-shoot-up-a-mosque-in-norway-was-inspired-by-christchurch-and-el-paso. Accessed: 26/02/26.

[Jaha23] Jahan, M.S. and Oussalah, M., 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546*, p.126232.

[Karr20] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. and Aila, T., 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8110-8119).

[KeTu24] Kelly, M. and Turton, W., 2024. Parler's New Owners Swear This Time Will Be Different. https://www.wired.com/story/parler-app-store-free-speech-censorship. Accessed: 26/02/26.

[Kine21] Kinetz, E. and Hinnant, L., 2021. Far-right networks harder to track as they go underground with cryptocurrency. The Times of Israel. 30 Sept. 2021. https://www.timesofisrael.com/far-right-networks-harder-to-track-as-they-go-underground-with-cryptocurrency/. Accessed: 26/02/26.

[Koeh19] Koehler, D., 2019. The Halle, Germany, Synagogue Attack and the Evolution of the Far-Right Terror Threat, CTC Sentinel, vol. 12, no. 11, Dec. 2019, https://ctc.westpoint.edu/halle-germany-synagogue-attack-evolution-far-right-terror-threat/. Accessed: 26/02/26.

[Kupp22] Kupper, J., Christensen, T. K., Wing, D., Hurt, M., Schumacher, M. and Meloy, R., 2022. The Contagion and Copycat Effect in Transnational Far-right Terrorism: An Analysis of Language Evidence. Perspectives on Terrorism, 16(4), 4–26. https://www.jstor.org/stable/27158149. Accessed: 26/02/26.

[Lako23] Lakomy, M., 2023. Artificial Intelligence as a Terrorism Enabler? Understanding the Potential Impact of Chatbots and Image Generators on Online Terrorist Activities, Studies in Conflict & Terrorism, DOI: 10.1080/1057610X.2023.2259195

[LeKo23] Lee, T. and Koltai, K., 2023. The Folly of DALL-E: How 4chan is Abusing Bing's New Image Model, Bellingcat, https://www.bellingcat.com/news/2023/10/06/the-folly-of-dall-e-how-4chan-is-abusing-bings-new-image-model/. Accessed: 26/02/26.

[Li25] Li, C., Wang, X., Li, M., Miao, B., Sun, P., Zhang, Y., Ji, X. and Zhu, Y., 2025. Bridging the gap between ideal and real-world evaluation: Benchmarking AI-generated image detection in challenging scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 20379-20389).

[LiWa25] Liu, F. and Wang, R., 2025. Fostering Parasocial Relationships with Virtual Influencers in the Uncanny Valley: Anthropomorphism, Autonomy, and a Multigroup Comparison. *Journal of Business Research*, vol. 186, 115024.

[LiZh25] Li, J., Zhang, J., Bai, X., Zheng, J., Zhou, J., Gu, L., 2025. InsTaG: Learning Personalized 3D Talking Head from Few-Second Video. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 10690-10700.

[LLAR] https://joshalosh.github.io/loki-docs/Lokinet/LLARP/. Accessed: 26/02/26.

Lu, W., 2024. Inevitable challenges of autonomy: Ethical concerns in personalized algorithmic decision-making. *Humanities and Social Sciences Communications*, 11, 1321. https://doi.org/10.1057/s41599-024-03864-y.

[Ma25] Ma, F., Yuan, Y., Xie, Y., Ren, H., Liu, I., He, Y., Ren, F., Yu, F. R., and Ni, S., 2025. Generative technology for human emotion recognition: A scoping review. *Information Fusion, 115*, 102753. https://doi.org/10.1016/j.inffus.2024.102753

[Mack19] Macklin, G., 2019. The Christchurch Attacks: Livestream Terror in the Viral Video Age, July 2019, Volume 12, Issue 6, https://ctc.westpoint.edu/christchurch-attacks-livestream-terror-viral-video-age/. Accessed: 26/02/26.

[Maia24] Maiano, L., Benova, A., Papa, L., Stockner, M., Marchetti, M., Convertino, G., Mazzoni, G. and Amerini, I., 2024. Human Versus Machine: A Comparative Analysis in Detecting Artificial Intelligence-Generated Images. *IEEE Security & Privacy*, *22*(3), pp.77-86.

[MaRi26] Mahara, A., and Rishe, N., 2026. Methods and trends in detecting AI-generated images: A comprehensive review. Computer Science Review, 60, 100908. https://doi.org/10.1016/j.cosrev.2026.100908.

[Mast] https://docs.joinmastodon.org/. Accessed: 27/02/26.

[Matz24] Matz, S.C., Teeny, J.D., Vaid, S.S. *et al.*, 2024. The potential of generative AI for personalized persuasion at scale. *Sci Rep* **14**, 4692 (2024). https://doi.org/10.1038/s41598-024-53755-0.

[Mirs20] Mirsky, Y. and Lee, W., 2020. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* 54, 1, Article 7 (December 2020), 41 pages. https://doi.org/10.1145/3425780.

[Mora25] Moravec, V., Hynek, N., Skare, M. *et al.* Algorithmic personalization: a study of knowledge gaps and digital media literacy. *Humanit Soc Sci Commun* **12**, 341. https://doi.org/10.1057/s41599-025-04593-6

[Mose25] Moseley, S., 2025. Automating Deception: AI's Evolving Role in Romance Fraud, CETaS Briefing Papers (April 2025).

[Naka08] Nakamoto, S., 2008. Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf. Accessed: 27/02/26.

[Nigh22] Nightingale, S.J. and Farid, H., 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, *119*(8), p.e2120481119.

[Nguy22] Nguyen, T.T., Nguyen, Q.V.H., Nguyen, D.T., Nguyen, D.T., Huynh-The, T., Nahavandi, S., Nguyen, T.T., Pham, Q.V. and Nguyen, C.M., 2022. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, *223*, p.103525.

[Pan23] Pan, B., Hirota, K., Jia, Z., and Dai, Y., 2023. A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing, 561*, 126866. https://doi.org/10.1016/j.neucom.2023.126866.

[Parl] https://www.reuters.com/article/business/google-suspends-parler-social-networking-app-from-play-store-apple-gives-24-hou-idUSL1N2JK05F. Accessed: 26/02/26.

[Pass24] Passos, L.A., Jodas, D., Costa, K.A., Souza Júnior, L.A., Rodrigues, D., Del Ser, J., Camacho, D. and Papa, J.P., 2024. A review of deep learning-based approaches for deepfake content detection. *Expert Systems*, *41*(8), p.e13570.

[Pawe22] Pawelec, M., 2022. Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions. *DISO* **1**, 19. https://doi.org/10.1007/s44206-022-00010-6

[PeLy21] Peters, J. and Lyons, K., 2021. Apple removes Parler from the App Store, The Verge. https://www.theverge.com/2021/1/9/22221730/apple-removes-suspends-bans-parler-app-store. Accessed: 26/02/26.

[Poco23] Pocol, A., Istead, L., Siu, S., Mokhtari, S. and Kodeiri, S., 2023. Seeing is No Longer Believing: A Survey on the State of Deepfakes, AI-Generated Humans, and Other Nonveridical Media. In: Sheng, B., Bi, L., Kim, J., Magnenat-Thalmann, N., Thalmann, D. (eds) Advances in Computer Graphics. CGI 2023. Lecture Notes in Computer Science, vol 14496. Springer, Cham. https://doi.org/10.1007/978-3-031-50072-5_34.

[PoDr16] Poon, J. and Dryja, T., 2016. The Bitcoin Lightning Network: Scalable off-chain instant payments. https://lightning.network/lightning-network-paper.pdf. Accessed: 27/02/26.

[Quin26] Quinn, B., 2026. Meet 'Amelia': the AI-generated British schoolgirl who is a far-right social media star. The Guardian. https://www.theguardian.com/politics/2026/jan/25/ai-generated-british-schoolgirl-becomes-far-right-social-media-meme? Accessed: 14/02/26.

[Rana22] Rana, M. S., Nobi, M. N., Murali B. and Sung, A. H., 2022. Deepfake Detection: A Systematic Literature Review, in *IEEE Access*, vol. 10, pp. 25494-25513, 2022, doi: 10.1109/ACCESS.2022.3154404

[Redd1] https://support.reddithelp.com/hc/en-us/articles/360043034412-What-is-a-Reddit-Premium-subscription#:~:text=Reddit%20Premium%20is%20a%20subscription,from%20the%20Reddit%20Premium%20page. Accessed: 26/02/26.

[Reut23] Reuters Fact Check, Video features deepfakes of Nancy Pelosi, Alexandria Ocasio-Cortez and Joe Biden. April 2023. https://www.reuters.com/article/fact-check/video-features-deepfakes-of-nancy-pelosi-alexandria-ocasio-cortez-and-joe-biden-idUSL1N36V2E0/. Accessed: 26/02/26.

[Ribe20] Ribeiro, M.H., Ottoni, R., West, R., Almeida, V.A. and Meira Jr, W., 2020. Auditing radicalization pathways on YouTube. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York, NY, USA: ACM; pp. 131-141.

[RoMaGo25] Romanishyn, A., Malytska, O., and Goncharuk, V., 2025. AI-driven disinformation: policy recommendations for democratic resilience. Front. Artif. Intell., 31 July 2025. Sec. Machine Learning and Artificial Intelligence, Volume 8. https://doi.org/10.3389/frai.2025.1569115.

[Roos18] Roose, K., 2018. On Gab, an Extremist-Friendly Site, Pittsburgh Shooting Suspect Aired His Hatred in Full, The New York Times, https://www.nytimes.com/2018/10/28/us/gab-robert-bowers-pittsburgh-synagogue-shootings.html. Accessed: 26/02/26.

[Sale22] Saleem, J., Islam R. and Kabir, M. A., 2022. The Anonymity of the Dark Web: A Survey, in *IEEE Access*, vol. 10, pp. 33628-33660, 2022, doi: 10.1109/ACCESS.2022.3161547

[Sang22] Sang, Y. and Stanton, J., 2022. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *International Conference on Information* (pp. 425-444). Cham: Springer International Publishing.

[Schr26] Schroeder et al. ,2026. How malicious AI swarms can threaten democracy. Science, vol. 391, pp. 354-357, doi:10.1126/science.adz1697.

[Seow22] Seow, J.W., Lim, M.K., Phan, R.C. and Liu, J.K., 2022. A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing*, *513*, pp.351-371.

[SiEdLe24] Simchon, A., Edwards, M., and Lewandowsky, S., 2024. The persuasive effects of political microtargeting in the age of generative artificial intelligence, *PNAS Nexus*, Volume 3, Issue 2, February 2024, pgae035, https://doi.org/10.1093/pnasnexus/pgae035.

[Sign1] https://signal.org/blog/signal-is-expensive/. Accessed: 26/02/26.

[Sign2] https://signal.org/docs/specifications/pqxdh/. Accessed 26/02/26.

[SimpleX] https://github.com/simplex-chat/simplexmq/blob/stable/protocol/overview-tjr.md. Accessed: 15/02/26.

[Sing19] Singh, S., 2019. Everything in Moderation An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content. *Open Technology Institute* (July 2019) 19 https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/. Accessed: 26/02/26.

[SMIDGE3.2] SMIDGE Deliverable D3.2 Horizon Scanning report 2016-2023 update.

[SMIDGE3.4] SMIDGE Deliverable D3.4 Report on Conspiracy, misinformation and extremist content update.

[SoTiRu25] Son, G., Tiemann, A. and Rubo, M., 2025. I am here with you: an examination of factors relating to social presence in social VR. *Frontiers in Virtual Reality*, 6, 1558233. https://doi.org/10.3389/frvir.2025.1558233.

[StLiAn22] Stein, J.-P., Linda Breves, P., and Anders, N., 2022. Parasocial interactions with real and virtual influencers: The role of perceived similarity and human-likeness. *New Media & Society*, *26*(6), 3433-3453. https://doi.org/10.1177/14614448221102900.

[Stoc24] Stockwell, S., 2024. AI-Enabled Influence Operations: Threat Analysis of the 2024 UK and European Elections, *CETaS Briefing Papers* (September 2024).

[Thal23] Thalen, M., 2023. 'Legitimate use of deepfakes': AI-generated video of Biden declaring World War 3, bringing back draft splits experts daily dot, Tech Posted on Mar 2, 2023, https://www.dailydot.com/debug/biden-deepfake-wwiii/. Accessed: 26/02/26.

[Thom20] Thomas, E., 2020. Manifestos, memetic mobilisation and the chan board in the Christchurch shooting, in: Counterterrorism yearbook 2020, Editor(s): Isaac Kfir, John Coyne, Australian Strategic Policy Institute (2020), URL: https://www.jstor.org/stable/resrep25133.7. Accessed: 26/02/26.

[Tolo20] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. and Ortega-Garcia, J., 2020. Deepfakes and beyond: A Survey of face manipulation and fake detection. Information Fusion, 64, 131–148. https://doi.org/10.1016/j.inffus.2020.06.014

[Tosi22] Tošić, A. and Vičič, J., 2022. Spatial Path Selection and Network Topology Optimisation in P2P Anonymous Routing Protocols. *Journal of Web Engineering*, *21*(1), pp.97-118.

[Vall23] Vallance, C., 2023. Facebook work filtering posts 'cost me my humanity', BBC News online, 25 April 2023, https://www.bbc.co.uk/news/technology-65346062. Accessed: 26/02/26.

[Wake22] Wakefield, J., 2022. Deepfake presidents used in Russia-Ukraine war, BBC, 18 March 2022, https://www.bbc.co.uk/news/technology-60780142. Accessed: 26/02/26.

[Wang25] Wang, H., Weng, Y., Li, Y., Guo, Z., Du, J., Niu, S., Ma, J., He, S., Wu, X., Hu, Q., Yin, B., Liu, C., Liu, S., 2025. Emotive Talk: expressive talking head generation through audio information decoupling and emotional

video diffusion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 26212-26221.

[WiPr24] Williamson, S. M., Prybutok, V., 2024. The era of artificial intelligence deception: Unraveling the complexities of false realities and emerging threats of misinformation. *Information*, *15*(6), 299. https://doi.org/10.3390/info15060299.

[WiPr24] Williamson, S. M., and Prybutok, V., 2024. The era of artificial intelligence deception: Unraveling the complexities of false realities and emerging threats of misinformation. *Information*, *15*(6), 299. https://doi.org/10.3390/info15060299.

[Woo25] Woo, S., Zubair, M., Lim, S. and Kim, D., 2025. Deep multimodal emotion recognition using modality-aware attention and proxy-based multimodal loss. *Internet of Things*, 31, 101562. https://doi.org/10.1016/j.iot.2025.101562.

[Xie26] Xie, S., Qiao, T., Li, S., Zhang, X., Zhou, J., and Feng, G., 2026. DeepFake detection in the AIGC era: A survey, benchmarks, and future perspectives. Information Fusion, 127, 103740. https://doi.org/10.1016/j.inffus.2025.103740.

[Yout1] https://support.google.com/youtube/answer/72857?hl=en-GB#:~:text=If%20you're%20in%20the,into%20the%20YouTube%20Partner%20Programme, Accessed:26/02/26.

[YXFL21] Yu, P., Xia, Z., Fei, J. and Lu, Y., 2021. A survey on deepfake video detection. *IET Biometrics*, *10*(6), pp. 607-624.

[Zhan22] Zhang, T., 2022. Deepfake generation and detection, a survey. *Multimed Tools Appl* **81**, 6259–6276. https://doi.org/10.1007/s11042-021-11733-y

[Zhen25] Zhen, D., Yin, S., Qin, S., Yi, H., Zhang, Z., Liu, S., Qi, G., Tao, M., 2025. Teller: Real-time streaming audio-driven portrait animation with autoregressive motion generation, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 21075-21085.