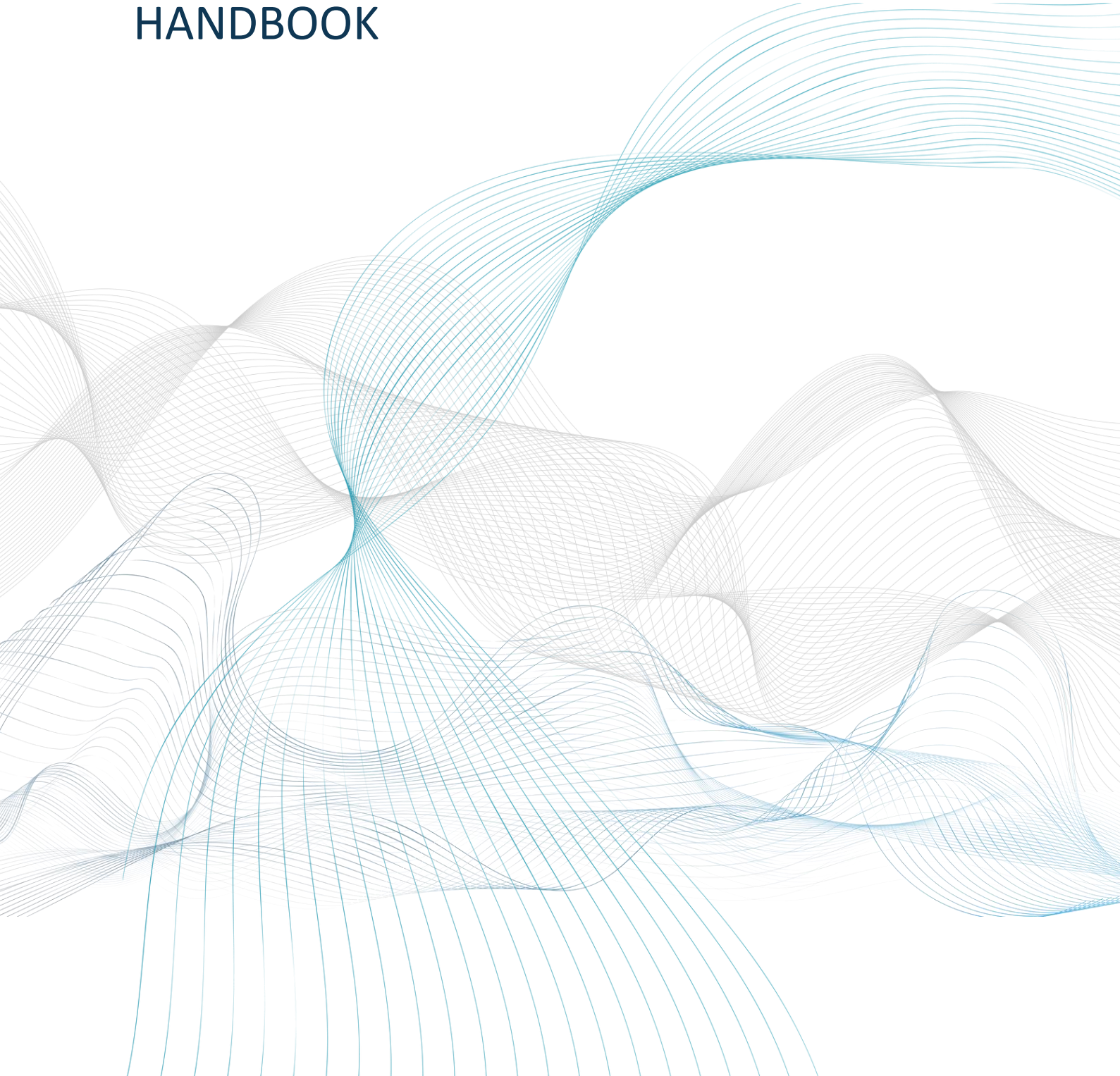
OppAttune | Countering Oppositional Political Extremism through Attuned Dialogue: Track, Attune, Limit

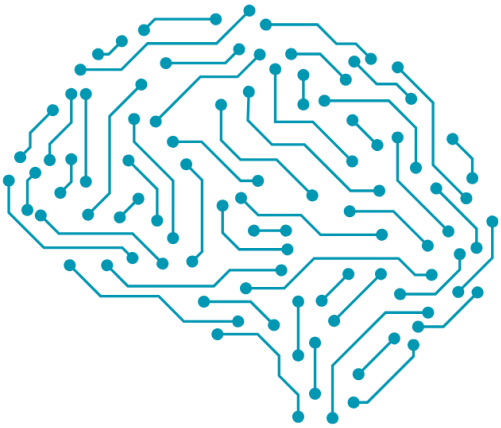
Tina Askanius / Despoina Filiou / Miriam Haselbacher / Griffin Rowell / Jullietta Stoencheva / Ursula Reeger

EVERYDAY EXTREMISM & ARTIFICIAL INTELLIGENCE

PUBLIC ENGAGEMENT PRACTITIONERS HANDBOOK



PUBLICATION DETAILS



Project Name:
OppAttune: Countering Oppositional Political Extremism Through Attuned Dialogue: Track, Attune, Limit

Grant Agreement ID:
101095170

Deliverable 4.6:
Public Engagement Handbook (Work Package 4 – Media, Machines & Mobilisations)

All illustrations and graphs are made using Canva

Funded by:

Co-funded by the European Union

Innovate UK

Contributing Partners:

MALMÖ UNIVERSITY

ÖAW ISR – INSTITUTE FOR URBAN AND REGIONAL RESEARCH

The Open University

Contact:
Tina Askanius, Malmö University,
tina.askanius@mau.se

Follow OppAttune Publications on:

oppattune.eu @oppattune oppattune @OppAttune @oppattune.project

TABLE OF CONTENT

INTRODUCTION

The aim of this *Public Engagement Practitioner Handbook* is to facilitate open, informed dialogue on contemporary challenges related to the intersection of technology and extremism, with particular attention to the ways in which current developments in artificial intelligence (AI) are shaping social and political discourse. The handbook provides a synthesis of the latest research on AI and online extremism, alongside key findings of the Horizon Europe project *OppAttune - Countering Oppositional Political Extremism Through Attuned Dialogue.* Together these insights serve to equip practitioners with perspectives and tools to engage diverse publics in critical reflections on the role of AI-driven media environments in both amplifying and mitigating the spread of extremist narratives in public discourse.

Artificial intelligence (AI) has rapidly emerged as one of the most consequential and contested technological developments of recent years. Developing at a rapidly accelerating pace, AI-driven systems increasingly mediate public discourse, shape everyday decision-making and structure our social interactions. Across Europe and beyond, policy- and decision-makers are grappling with the challenges of regulating and governing these systems as well as with questions of political responsibility of their impact. EU regulatory initiatives such as the Digital Services Act (DSA) and the Artificial Intelligence Act (AIA) signal a clear ambition to establish stronger regulatory frameworks, enhance transparency and set higher standards of accountability for tech platforms and AI technologies. The DSA - adopted in 2022 and fully applicable in 2024, has introduced new rules for platform transparency, user protection and systemic risk mitigation. The AIA followed in 2024 as the first comprehensive AI regulation designed to govern the AI systems behind moderation and recommendation based on the protection of fundamental rights and freedoms. However, the practical interpretation, long-term impact, and effective enforcement of these laws remain uncertain. At the centre of current debates over compliance and enforcement lies the challenge of balancing digital security, economic competitiveness, and innovation with the protection of fundamental rights and freedoms. At the moment, there is no clear consensus on AI governance, leaving gaps, tensions, and opportunities for developers, policymakers and platforms to shape the development but also for actors involved in the production and circulation of extremist content to exploit regulatory ambiguity. These dynamics are particularly salient in efforts to address online extremism and the evolution of political extremism more broadly. This handbook contributes to the broader efforts of the OppAttune project to design inclusive and participatory ways of sharing knowledge beyond academia addressing these pressing challenges. It aims to identify and synthetise key insights from existing scholarship that can guide future public engagement initiatives, particularly for NGOs and civil society actors working in the field of radicalisation and extremism prevention and democratic resilience.

OppAttune tracks the evolution of oppositional ideologies and aims to develop tools to track, attune and limit the spread of extreme political narratives. The project operates at micro, meso, and macro levels, comprehensively addressing these issues from the perspective of legal, political, and historical factors to group dynamics, media contexts and the everyday life of citizens. The following work contributes directly to OppAttune's broader goal of limiting the evolution of extreme political narratives by attuning dialogue and providing strategies for awareness-raising and public engagement, around these urgent challenges threatening inclusive and cohesive European societies. Within this larger framework, Work Package 4 *Media, Machines & Mobilisations* brings together an interdisciplinary research team to identify and understand contemporary extremist narratives on- and offline and to map their spread across the digital mainstream and into local ecologies across Europe. The handbook synthesises key insights from the empirical research conducted within WP4 on the circulation and mainstreaming of extremist narratives in online spaces - what we refer to as *everyday extremism.*

Everyday extremism is usefully understood as recurrent and normalised presence of extremist narratives embedded in familiar discourse, spaces, and practices (see Askanius et al., 2024a; 2024b; Andreouli et al., 2024; Rottmann et al., 2025). It pertains to mundane discourses and practices rather than discourse and strategic communication produced by extremist actors or around extraordinary events, and tends to be amorphous, ambiguous, and embedded in the routine aspects of daily life. Everyday extremism highlights how extremist narratives – including violent, exclusionary and dehumanising ideas - increasingly penetrate everyday settings influencing ordinary European citizens' daily social interactions, media consumption, and public discussions. Recognising these dynamics is essential for understanding how extremist worldviews become part of the cultural mainstream and for developing strategies to counter its influence.

"Everyday extremism refers to the recurrent and normalised presence of extremist narratives embedded in familiar discourse, spaces, and practices. It pertains to mundane discourse and practices rather than extreme actors or extraordinary events, and tend to be amorphous, ambiguous, and embedded in the routine aspects of daily life."

(Askanius et al., 2024)

In this context, particular attention must be paid to content that aligns with PRIME dynamics—prestigious, moralized, emotionally charged, and linked to ingroup identity—which research shows provide especially effective pathways through which violent and exclusionary ideas gain visibility and traction in everyday, seemingly innocuous contexts (Brady et al., 2023). Everyday extremism emerges through subtle cues, emotional triggers, coded language, and subtle references that circulate in mainstream digital spaces. The emergence of new and continually shifting ways in which PRIME-aligned content can fuel extremist narratives and actions requires public and private actors to respond with comprehensive strategies taking into account key lessons and evidence from both computer and social science perspectives.

Written at the conclusion of the OppAttune project, this handbook provides an overview of recent scholarship on the impacts of AI, automation, and platformization of political discourse and everyday civic life online, highlighting the role played by these dynamics in the penetration of violent ideas and political extremism into these domains. It concludes by offering recommendations for how to engage the wider public in conversations about the polarising and dehumanising ideas that increasingly shape everyday social and media environments, and the role of technology in these processes. It encourages practitioners to consider both the risks and opportunities of AI technologies - how they may amplify divisions, but also how they can be leveraged to support inclusion, critical reflection, and democratic participation. On doing so, the handbook seeks to contribute to the development of informed and reflective publics capable of recognising and resisting everyday extremism.

We start by outlining key concepts and terms that are crucial for understanding current developments in AI, before introducing the work we have conducted within the OppAttune project. We then provide a brief overview of the most recent research in this rapidly evolving field, focusing on the role of AI, automation, and algorithmic systems in amplifying, circulating, and potentially limiting extremist narratives online. As such, the handbook connects current debates about technological developments with findings from WP4's empirical analysis of how extremist ideas circulate across Europe's digital and social landscapes. By bringing these bodies of knowledge into dialogue, the handbook identifies key lessons from existing research and perspectives from citizen science projects to inform engagement strategies that can support practitioners in fostering public awareness and resilience on these critical issues.

KEY CONCEPTS AND DEBATES

This handbook is reliant on a number of key concepts and debates central to the latest scholarship on the role of advanced technology on extremism. These are all "big" contentious concepts which themselves are susceptible to politicization and distortion, and therefore subject to intense academic and political debate, as made evident from the discussions below. Using systems thinking to guide our classification of model typologies, in this section and throughout the handbook we discuss three main AI models whose end goals are to generate content, make decisions, and recognize patterns (Baele & Brace, 2024). Of these three, algorithms as understood and used in this handbook feature most prominently in the latter two models. In this section, we briefly outline the central issues animating critical debates surrounding the most recent developments in this area.

DEBATES ON THE BUILT-IN BIAS OF LARGE LANGUAGE MODELS

There is nothing inherently neutral about the large language models (LLMs) and other generative AI systems that produce text, images, and audio, nor about the platform-level algorithms that determine how such content is ranked, circulated, and made visible to users. On the contrary, AI technology has emerged out of racially-charged and inequitable political contexts significantly influencing the ways in which the technology developed and used (Burton 2023). Trained primarily on large-scale, human-generated datasets, and increasingly on synthetic, computer-generated content - AI systems tend to reproduce and, in some cases, intensify the biases, prejudices, political assumptions, and value hierarchies embedded in the societies that produce them (see e.g. Benjamin 2018; Eubanks 2018; Noble 2018).

While platforms have implemented filters to detect and remove certain forms of violent or explicitly extremist language, far less effort has been devoted to understanding, categorising, and mitigating the coded or more subtle discursive forms through which extremist ideas circulate and gain legitimacy. Even under ideal conditions – including the presence of large, well-resourced, and well-trained human moderation teams - platforms would still face substantial challenges in addressing these dynamics. In practice, however, many human content moderation roles have been eliminated in recent years in favour of automated systems (Corella & Moreno 2025). Those moderators who remain are often constrained by narrow and inconsistently applied hate speech and violence guidelines, typically developed with limited input from the marginalised communities they are ostensibly intended to protect.

As a result, contemporary systems of content moderation and algorithmic amplification routinely fail to address overt racism, classism, sexism, homophobia, violent threats, and other forms of extremist discourse. These failures should not be understood simply as technical shortcomings. Rather, they reflect governance choices and incentive structures that enable - and in some cases actively reward - the circulation of polarising and extremist content. This problem has been exacerbated by an overreliance on automation that lacks a grounding in critical social science research and insufficient engagement with the social and cultural contexts in which these systems operate. It has also been intensified by broader political, economic, and social shifts, which existing scholarship suggests digital technologies significantly contribute to, even if they are not the sole driving force.

STRUCTURAL BIASES IN AI AND ALGORITHMS

AI tools, the data they draw upon, and the code that underpins them are produced and funded within existing systems of oppression, discrimination, and exclusion. As a result, and as shown by recent scholarship, these technologies often exacerbate the very social and material inequalities from which they emerge (Williams 2024; Corella and Moreno 2025). Social prejudices that manifest in both extremist and violent rhetoric and actions are therefore reproduced and embedded within technological systems, particularly in the absence of adequate oversight, representation, and care in their design and governance. Large language models (LLMs) and other technologies based on machine learning are trained on publicly available datasets, including word association studies that frequently reveal deep seated human prejudices and, in some cases, extremist views (Corella and Moreno 2025; Wyer and Black 2025). These data have not always been carefully curated, and critical contextual information is often missing, both at the level of model training and in the subsequent use of these systems to generate or circulate content shaped by such prejudices (Williams 2024).

AI and algorithmic content moderation systems are trained on largely the same data that produces extremist content and discourses (Oh & Downey, 2024; Corella & Moreno, 2025). It would be impossible to train these technologies on entirely neutral content, such that it exists, as it would be even more woefully equipped to identify and categorize content as extreme. Even seemingly innocuous training data or information such as postal codes is representative of characteristics, trends, and indicators of human social systems that contain biases that are the result of biases, histories, and prejudices, some of which are discriminatory or violent (Corella & Moreno, 2025). These forms of discrimination and bias inevitably lead to inequality and can undermine human rights when AI systems are not thoughtfully designed and responsibly governed (Solomon & Spanò, 2023).

Extremist actors with resources, including politicians, states, and hate groups, can take advantage of the quantity-over-quality approach taken by many tech platforms responsible for both open and closed source LLMs and generative audio and visual products. The economic bar is to create and influence online content is not prohibitive for determined groups and individuals (Cleveland-Stout, 2025). Because of the lack of database curation, particularly carried out by humans, used by many popular tech companies, these actors are able to flood the internet with content (Baele & Brace, 2024; Maynor, 2025).

Scholarship notes that this can be particularly overwhelming to algorithms and human users alike due to the scale, velocity, and precision in which synesthetic content can mimic human speech or content (Baele & Brace, 2024; Gandhi, 2024). This has a number of downstream social and technological effects, including increasing the ability of users looking to AI platforms to create extremist content, or content that serves as ‘evidence’ for extremist narratives, shifting the overall data pool and subsequently algorithms in a way that, by-default, skews the information, answers, and perspectives towards extremist content and ultimately their overall agendas and narratives (Baele & Brace, 2024; Maynor, 2025).

RECOMMENDATION ALGORITHMS AND FEEDBACK LOOPS

Recommendation algorithms used by social media platforms such as Tik Tok, Facebook, and YouTube are deeply embedded in the user experience of platforms and are difficult, cumbersome, or impossible to opt out of. The more provocative or sensational the content, which often features components of PRIME, the more it keeps users involved and the more content of a similar nature the algorithm suggests (Shin & Jitkajornwanich, 2024). Even for viewers who are new to a platform or are otherwise unrecognized by the system, algorithms have been shown to suggest content that is extreme or hateful. Furthermore, existing platforms users who do not actively engage with or seek out extreme content still have it recommended to them and places in areas of the platform that recommend content, such as on its "For You Page." Extreme content is given preferential treatment in the algorithm, leading to violent content being a reliable path to going viral, generating revenue for both platform and poster as more content is shared and viewed. Because the biased algorithms train themselves on data from users who have previously been subjected to and engaged with the algorithm, a feedback loop is generated (Shin & Jitkajornwanich, 2024).

Research suggests that the personalization of these algorithms, driven by user engagement and the data it creates, is even stronger for populations with extreme ideological beliefs. This results in "rabbit hole" effects, further radicalizing those with extremist worldviews, including those sharing content that encourages physical violence (Shin & Jitkajornwanich, 2024). While the psychological and social processes of radicalization take time, algorithms work fast to suggest content of a similar nature. YouTube and Tik Tok's algorithms are documented as recommending up accounts with far-right views even after minimal user engagement with the platform (Shin & Jitkajornwanich, 2024). Once users have viewed or engaged in this content, alternative voices from the political and social spectrum are recommended significantly less frequently.

The focus of many of the major large language models (LLMs) has been quantity over quality. Not only is there a lack of carefully curated datasets with which to train LLMs on, there has been a systemic overall lack of consideration of the biases and differences in representation and voice that exist in the open internet (Wyer & Black, 2025). This has helped create a feedback loop in which certain voices are overrepresented or underrepresented, creating an artificial hegemony that can bring extreme views from the fringe into the mainstream, normalizing them (Wyer & Black, 2025). Because algorithms currently used to detect effectively extreme content are largely ineffective, and because extremist content benefits from algorithmic support, extremist views can quickly become hegemonic in certain corners of the internet–especially when platforms do not sufficiently limit bots that reproduce similarly extreme content. While Regulation (EU) 2024/1689, the West's most expansive AI regulation, comes into effect next year, it remains to be seen how effective this legislation will be when it comes to compelling tech companies to label AI content. Literature indicates that the mad dash towards AI competitiveness and adoption, including by many European countries, has been done without substantial user and public input, posing significant risks for marginalized communities.

DEBATES ON ALGORITHMIC RADICALISATION

Automated algorithms not only connect isolated individuals but can also play a role in radicalizing individuals (Barnes, 2022). For example, research suggest that users who seek out and watch a video on YouTube will autoplay recommended videos the algorithm deems as similar by default, regardless of whether the previous video shares extremist views or content (Barnes, 2022). Because of the volume of users that view extremist content, whether they want to or not, and are subsequently recommended more, even a fraction of viewers who are demonstrated as becoming radicalized in large part to the extremist content they are exposed to presents a significant issue. Socially isolated or otherwise vulnerable viewers of extremist content are at higher risk of being radicalized or further radicalized by content pushed or otherwise made accessible to them by algorithms because the content can reinforce beliefs, provide reassuring solutions, and community. Inflammatory content that otherwise would be banned often isn't due to specific, technical reasons in the case of AI or due

to the subjectivity of human content moderators (Barnes, 2022). Even when users break platform rules for hate speech or other prohibited content, platforms rarely remove them from the services due to the business model of the platforms which is reliant on engagement to generate revenue. This is an issue that is difficult to address given the obscured, opaque "black box"—the hidden, internal machinations of AI systems—nature of many major algorithms. A lack of shared platform guidelines and enforcement mechanisms regarding extreme, violent, and hateful content has created too many pathways for users, assisted by functionally unregulated AI companies, to flood the internet with harmful, dangerous content.

Recommendation algorithms used by major platforms bring people to extremist hate groups, sections of platforms that can communicate directly and exclusively with those in the group, on platforms even when users are not looking for them (Barnes, 2022). Experts agree that until algorithms are trained to not optimize engagement above everything else, platforms and the technologies they host will continue to accelerate extremism. Users who share and engage with extremist content are more likely to be recommended to users who do not share those beliefs, as this also drives platform engagement (Barnes, 2022). Using automated methods alone, including AI tools, has not yet been demonstrated to be an effective tool to enforce content moderation.

Taken together, the processes of algorithmic radicalization foreground the structural limits of content moderation systems that operate within the same engagement-driven, opaque platform logics that enable extremist content to circulate and persist in the first place. Studies have shown that changes to algorithmic moderation alone, largely supported or proposed by tech platforms, are ineffective at significantly altering the habits or views of those who hold extremist views. Such solutions downplay or minimize the still substantial role that humans play in AI and algorithmic content moderation. Such work often operates under poor conditions and exposes workers, often from the Global South, to significant psychological harm. Even the most recent literature notes that AI and algorithms are still largely dependent on human workers to train computer models. The scale of content on the internet has far outpaced the ability to train workers, many from the Global South who work under poor conditions, effectively. Companies have also scaled up the use of automated and AI systems of content moderation, while scaling back on the critical part of human workers.

INSIGHTS FROM THE OPPATTUNE PROJECT

Within the OppAttune project, work Package 4 “Media, Machines and Mobilisations” focuses on understanding extremist narratives at the levels of content, and the mechanisms through which they circulate between online and offline contexts. This follows the assumption that online dynamics such as incitement to violence, divisive and exclusionary language can translate into offline violence and exclusionary behaviours (see also Solomon & Spanò, 2023). While this work has focused primarily on how extremist narratives spread through visual ephemera online and the interplay between fringe and mainstream digital platforms, it has been less concerned with the broader socio-technical infrastructures and politics of tech regulation that shape the nature and circulation of these narratives. This includes algorithmic and platform logics, business models, and recent breakthroughs in artificial intelligence, which has advanced rapidly since the OppAttune project began. The handbook brings the work of OppAttune into dialogue with recent developments in these areas.

Work package 4 focused on the immediate everyday environments in which people live and interact in a digitalized and globalized world which unfolds across online and offline spaces. The concept of the everyday has thus been an important analytical lens and entry point for our studies, centering our studies on the ordinary, habitual aspects of daily life, mundane experiences and practices as well as interactions and conversations between citizens that happen on a regular basis. Drawing on social theory, we consider the everyday not as a trivial backdrop, but as a site where social meanings, power structures, and identities are reproduced and negotiated on a daily basis. In doing so, the work package situates extremism within the broader fabric of everyday life in a digitalised society.

Processes of digitalisation have fundamentally reconfigured everyday interactions, routines, and communicative practices relocating them within the digital realm. Activities that once took place in physical spaces - and continue to do so - such as socialising, working, shopping, or accessing information, now also take place online and are mediated by digital platforms and AI technologies. This transformation has changed how people connect and communicate, while also reshaping our sense of space, time, and social belonging. The boundaries between online and offline life are increasingly blurred and digital practices and discourses have become deeply intertwined with physical and social realities.

Within our research, we therefore focused on everyday online spaces, e.g. the most popular social media platforms, discussion forums, and content-sharing platforms that people routinely inhabit, to examine how users encounter and interpret extremist discourses. This is a critical area in which citizen science approaches can be integrated into efforts around curbing AI-driven extremism. NGOs can empower citizen scientists to, at different levels, identify the ways in which extremism exacerbated by technology is observed or not observed in people’s everyday digital and non-digital lives. The

participatory model of citizen science makes it possible to gather meaningful, context-specific local data—information that researchers working outside underrepresented communities often struggle to access. Because AI systems and algorithms rely on human labour and inevitably reflect human biases and prejudices, it is crucial that solutions are co-created with citizens from all backgrounds, especially those most targeted by extremist ideologies. The responsibility for identifying problems and proposing solutions should not fall on any single group. Instead, communities working together can improve the quality and depth of data related to the issues addressed in this handbook and help reimagine what ethical AI and algorithmic systems should look like (Fraisl et al. 2025).

By observing how extremist narratives appear within memes, comment sections, news feeds or otherwise algorithmically suggested content, we sought to understand how users encounter such narratives within the flow of everyday online interactions. The following sections presents key insights from our research conducted in Austria, Bulgaria, and Sweden that are relevant to this handbook, drawing together findings from two research reports, the *Visualisation report of emerging extremist narratives across Europe* (Askanius et al., 2024a) and the report on *Translocal articulations of extremism in three countries* (Askanius et al., 2024b).

UNDERSTANDING EXTREMIST NARRATIVES

Extremism is a complex phenomenon that goes beyond isolated acts of violence or worldviews. Key to understanding the discursive powers of extremism are the narratives used to justify actions and worldviews, by framing social conflicts in terms of rigid and divisive us-versus-them distinctions. Research highlights how extremist narratives, understood here as violent, exclusionary and dehumanising framings of specific social groups, mobilise support for and legitimise violence or hostility towards perceived out groups in ways that breach basic democratic principles of human equality and dignity.

> ""[E]xtremist narratives can be understood as interconnected systems of stories that promote the idea that the success or survival of the in-group depends upon taking hostile or violent action against the out-group"
>
> (Berger 2018)

This work builds on a definition of extremism grounded in social identity theory (SIT). From this perspective extremism is understood not as a fixed point but as a spectrum of beliefs, emerging from contentions identity dynamics between constructed in-groups and out-groups (Berger 2018). While extremist movements and groups are heterogenous in their organisation and agendas, Berger (2018) argues that all share a common value proposition: the belief that the extremist group (the in-group) offers a solution to a perceived crisis caused by a perceived threat from an opposing out group. Extending this argument, extremist narratives can be understood as interconnected systems of stories that promote the idea that the success or survival of the in-group depends upon taking hostile or violent action against the out-group (Berger 2018).

Even before explicit extremist ideas become visible, actors'—influential and otherwise—utilization of PRIME characteristics results in content that is deceptively ordinary and difficult to dislodge from mainstream political discourses, an issue made worse by the use generative content that both shapes and is shaped by existing social dynamics in data. However, this framework allows for violent content—as expressed through language and visual forms of communication, like images, symbols and emojis—to be analysed in a systematic manner. This also pays tribute to taking an "everyday" approach to extremism as it allows to examine how extremist narratives circulate in ordinary everyday contexts—be it on stickers in the streets or memes online—rather than focusing solely on overtly violent statements or acts in the context of terrorism and violent extremism. This perspective takes

into account how seemingly small expressions of hostility, discrimination and violence contribute to a broader climate of polarization and exclusion based on the construction of in- and outgroups.

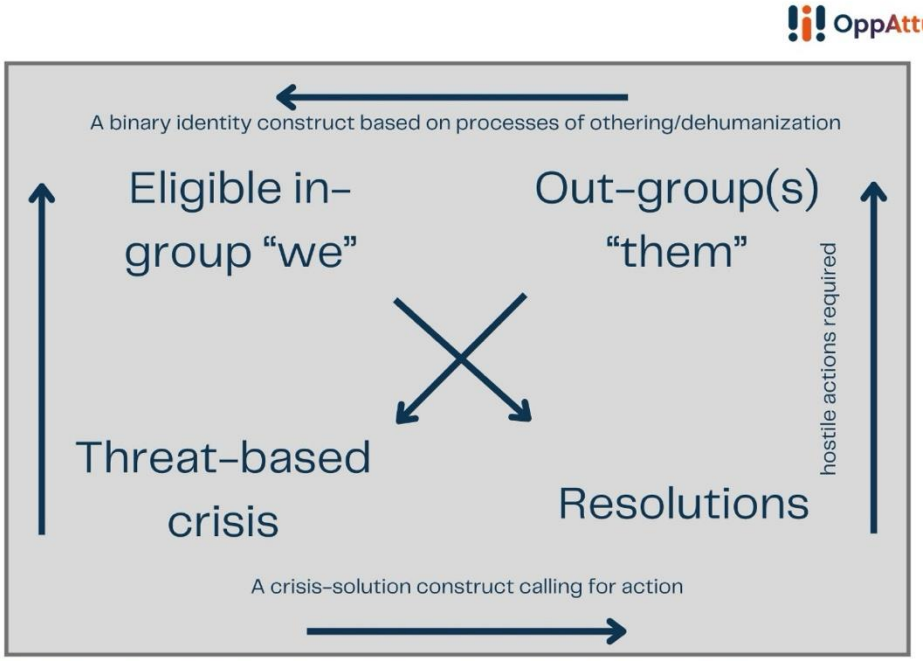


Fig. 1: The value proposition of extremist narratives, adapted of Berger 2018

From this follows that hostile or violent actions do not necessarily need to be expressed as or involve direct physical harm. Instead, violence can take on structural and symbolic forms as well. In our empirical research, we have demonstrated how these forms of violence exist on a continuum and are closely interconnected. Symbolic violence refers to actions or representations that dehumanize certain groups, typically manifested through sexist, misogynistic, racist, or anti- "woke" content. Structural violence, by contrast, occurs by targeting political or institutional mechanisms, such as conspiracy narratives or narratives directed at a specific elite that is accused of abusing its powers. Physical violence appears in explicit references to causing physical harm or otherwise eliminating out-groups. Together, these different forms of violence interact and reinforce one another, revealing how everyday extremism can escalate into more overt forms of violent threats.

AI technologies, particularly recommendation systems and content filters, influence how extremist narratives circulate online by amplifying divisive content and reinforcing patterns of engagement. Extremist narratives have the potential to spread rapidly and widely online due to the speed and reach of digital platforms. Social media algorithms often amplify polarizing content, allowing hostile messages to gain visibility and resonance across communities. The anonymity and distance of online spaces further lower social barriers, making it easier for users to express or endorse violent and hateful ideas without facing immediate consequences. As a result, extremist discourses circulate, normalize, and intensify far more quickly than in offline settings.

FROM THE FRINGE TO THE MAINSTREAM AND VICE VERSA

Research on online radicalization and extremism often distinguishes between so-called fringe and mainstream platforms (Conway, 2017; Ebner, 2023; Munk, 2024; Zhang & Davis, 2024), even if recognising that the two cannot always be neatly separated. Not least because platform architecture, governance regimes and engagement incentives can rapidly change as illustrated by Elon Musk's recent takeover of Twitter/X which reconfigured moderation practices and content visibility of the platform. Fringe platforms, such as Rumble, BitChute, Odysee, Telegram, or Discord provide relatively unmoderated environments that foster anonymity, community-building, and often unhindered circulation of extremist content. In contrast, the household names in the social media and content

industry, reach wider audiences, are easy to access and have community guidelines as well as content moderation that allow for de-platforming and the erasure of content. However, obscured to algorithms, digital capitalism and monopolization, and political changes at the level of the U.S. federal government mean that these mainstream platforms are themselves breeding grounds for extremist content that is organic, if only briefly, to them in addition to hosting and spreading extremist content created on non-ubiquitous platforms. Extremist actors that previously operated in the margins of the internet are now uniquely positioned to shape global conversations in ways that are becoming impossible to ignore. Our research illustrates how popular local discussion forums like Sweden's Flashback Forum or Bulgaria's BG-Mamma can become conduits for extremist narratives. Many of the narratives that transition from fringe spaces into mainstream visibility display PRIME characteristics, providing opportunities for literacy, education, and engagement.

While de-platforming efforts have pushed many extremist actors toward fringe platforms, these spaces continue to amplify their messages, offering anonymity and limited moderation. The circulation between mainstream and fringe platforms through users, as they cross-post, remix, and comment, ensures that extremist narratives remain visible and easily adaptable. This interconnectedness of online platforms creates a dynamic environment and the fluid movement between digital space demonstrates that the relationship between fringe and mainstream is not linear but recursive: extremist narratives depend on mainstream exposure to gain traction. At the same time, platforms are depicted as enemies due to content moderation and de-platforming, which further fuels the narrative of resistance and persecution, and which supports fringe platforms. This dynamic creates a cycle where mainstream visibility is leveraged, while the mainstream is simultaneously positioned as an oppressive force. Together, these dynamics reveal a complex ecosystem in which extremist ideas, although presented differently, flow across digital boundaries, gradually embedding themselves in the ordinary, everyday experiences of online life.

Underlying these dynamics are the algorithms that shape online visibility and engagement. As the section on recommendation algorithms and feedback loops has outlined, social media algorithms encourage content that provokes strong emotional reactions, like anger, fear, or outrage, which often align with the affective tone of extremist discourse. As users interact with provocative material, recommendation systems amplify similar content, reinforcing ideological echo chambers and contributing to individual radicalisation processes. Widening and deepening these echo chambers has become useful for social media platforms as algorithms push fringe content into the mainstream with little context, restriction, or warning because exposure to strong narratives that draw on PRIME content or other forms of sensational content that drive engagement of all kinds.

HUMOUR AND VISUAL DIGITAL CULTURES IN THE CIRCULATION OF EXTREMIST NARRATIVES

A starting point for our empirical research is the observation that humour, irony, and ambiguity characteristic of contemporary digital cultures can function as Trojan horses, disguising violent, exclusionary and dehumanising ideas in layers of pop-cultural familiarity and digital playfulness. Humorous and strategically ambiguous digital formats such as memes have long been used as a means of masking or softening controversial and harmful ideas. Extremist actors have been documented to employ satiric forms and irony to cloak their messages, allowing them to appeal to a wider audience without immediately triggering backlash. Explicit hate speech and violent threats are obfuscated using of popular and subtle visual cues, memes, emojis, satirical content, and pop cultural references, that obscure the underlying violent and dehumanizing messages. Far-right actors too strategically exploit such aesthetics and communicative registers to attract younger audiences, building credibility through in-group language and countercultural rebellion of digital subcultures (Goetz et al., 2018; Lewis, 2018). This technique renders such content seemingly innocuous, increasing its circulation across mainstream digital spaces and reducing the likelihood of immediate detection or moderation.

For example, within the data we analysed, emojis were frequently used as subtle yet powerful symbols carrying (coded) messages targeting certain out-groups. For example, an airplane emoji is frequently

used to call for the deportation of immigrants, invoking the idea of forced removal or exclusion without explicitly stating it. Similarly, national flags are frequently employed in digital spaces as symbols of nationalism, often to reinforce exclusionary or supremacist messages, while blending seamlessly into everyday online conversations, subtly promoting divisive or violent political agendas. Memes also play a central role in this process. They rely on context-dependent humour and ambiguity to convey messages. Memes are quick to produce, easy to share, and highly accessible making them a particularly effective format for spreading content rapidly. They often blend satirical commentary with serious political messages, normalising extremist rhetoric and making it easier for such ideas to infiltrate mainstream political discussions.

The translocal nature of memes, emojis and other elements of visual cultures further strengthens the possibility to spread extremist content, as they are easily understood across local and national contexts. This adaptability makes them a potent tool for spreading extremist ideologies across a range of political and social contexts. The use of provocative, controversial, or well-known imagery, such as symbolic references to violence or subversive political themes or actors, can tap into a sense of collective identity, reinforcing in-group cohesion while simultaneously demonizing and dehumanising perceived enemies (see Stoencheva et al 2026).

Moreover, the use of humour and irony in memes and other forms of online content conveying extremist narratives provides a layer of plausible deniability. Plausible deniability complicates efforts by platforms and authorities to identify and moderate harmful content. Violent and otherwise harmful messages are easily written off as simply jokes, irony or hyperbole complicating efforts to call it out as well as the enforcement of content moderation policies. Cloaked and coded messaging is another key challenge. For example, the use of terms like “culture enricher,” a coded expression once associated with neo-Nazism is reappropriated in Sweden to refer more broadly to immigrants, allowing hate speech and racist discourse to circulate freely without immediate detection or repercussions (Åkerlund 2021).

These empirical insights from the OppAttune project on the aesthetics and circulation of contemporary forms of extremism informs the following discussions of existing research on AI and its implications for political discourse across and beyond social media platforms. These insights also help shape our recommendations for incorporating citizen science perspectives into public engagement projects, highlighting approaches that empower communities to critically engage with algorithmic systems and participate in efforts to mitigate the harmful effects of online extremism.

KEY INSIGHTS ON AI AND EXTREMISM

Emerging evidence on the intersections between AI and extremism draws from diverse fields, including computer science, digital media studies, critical algorithmic studies, extremism and terrorism studies and social psychology. While research bringing together results from these diverse fields remains scarce, recent studies have begun to connect perspectives to explore how AI technologies can both drive and combat extremist ideologies (Burton 2023; Gal et al., 2022; Solomon & Spanò, 2023).

Crudely, existing research suggests that while AI is often framed as a tool for detecting and countering extremism, AI and algorithmically driven platforms can also intensify extremist dynamics, by amplifying harmful content, enabling the dehumanisation and targeting of out-groups, and contributing to political violence and societal polarisation (Burton 2023). Importantly the literature emphasizes that AI and algorithmic systems do not generate extremist beliefs or behaviors on their own, rather they shape the visibility, scale and circulation of extremist content online thereby contributing to its growth. Overall, scholars agree that challenges related to AI and extremism should not be understood as purely technological problems, but as deeply political and economic issues tied to platform incentives, power asymmetries, and broader political contexts influencing development and regulation.

As a consequence, attention has increasingly shifted from abstract debates about AI technologies to the concrete governance practices of social media platforms and the ranking, recommendation and engagement optimisation algorithms they rely on. One major issue identified by scholarship is the lack of economic incentives for social media companies to meaningfully limit divisive and inflammatory content, even when such content is synthetic or generated by bots. Social media platforms generally do not use comprehensive human and AI-monitored guardrails that mark or otherwise limit AI-generated content, including computer generated text shared by bots (Aljabri et al., 2023). Platforms have economic incentives that are at cross purposes when it comes to limiting these bots, including and at times especially, ones that produce extremist content because up to certain points they also drive engagement (Delkhosh et al., 2023). Additionally, these systems and the significant amount of human labour involved in content moderation, struggles to identify extremist human-produced content that reproduces or iterates on content that is created entirely by a computer.

Against this backdrop, scholars have noted that open-source Text-To-Image systems (TTI) increasingly play a major role in shaping AI-generated visual and auditory content. Overwhelmingly, the content created and influenced by these systems disproportionately feature inflammatory and extreme content that skews toward the male gaze, objectification, violence, and harmful, reductive stereotypes, not least of women. This material does not remain confined to open-source environments; rather, it feeds back into the broader AI ecosystem, influencing the training data and

outputs of closed TTI systems as well (Wagner & Cetinic, 2025). Exploitative content quickly spreads on platforms due to algorithms that reward divisive content, outrage and conflict. These dynamics are further intensified by recent advances in diffusion models, i.e. systems that iteratively remove noise before reversing the process to generate highly realistic outputs from user prompts. Such models enable increasingly granular, high-quality modifications, including the use of carefully calibrated language that elicits extreme or exploitative imagery without formally violating platform policies. As a result, automated moderation mechanisms face growing difficulty in detecting and regulating harmful content. (Wagner & Cetinic, 2025).

Further, generative audio produced using platforms such as *Udio* and *Suno* has been used by extremist actors to quickly create music that spreads violent and discriminatory rhetoric (Lopes, 2024; Tencer, 2024). One method taken to skirt the loosely enforced rules that govern audio content is to generate the text on other platforms. Generative audio can also mimic humans effectively, leading to instances where well-known public figures' voices are used to create extreme or misleading content (Fisher, Howard, & Kira, 2024). It is difficult for both algorithms and users to distinguish authentic audio from synthetically generated material, due in part to the limited adaptation of robust watermarking techniques and the relative ease with which existing automated audio watermarks can be altered or removed.

MALIGN ACTORS LEVERAGING AI – WHAT DO WE KNOW?

Extremist actors—ranging from religious to white supremacist actors—increasingly take advantage of computer-generated tools to create and spread their message (Verma, 2024; Tondo, 2025). Even unwitting users are still able to produce extremist content by using text-based commands that do not use bigoted or extreme language (Górska & Brzeziński, 2025). Scholarship notes there is a careful line that extremist actors leveraging tech platforms walk in order to grow their influence (Allchorn, 2024). Extremist actors who are vocally sceptical or critical of AI and share extremist views on the topic using social media platforms are largely tolerated, given its ability to generate engagement (Awasthi, 2024; Shin, 2024). Extremist actors, and the bots they create, are adept at adopting "algospeak," coded language and characters that are effective at evading automated content moderation, which is not yet good at removing iterative visual posts, or catching new algospeak (Golunova, 2025).

Extremist actors are also able to leverage improvements in speed and accuracy via LLM and generative audio/visual models to content in an ever-increasing range of languages, including using voices and likenesses from prominent people who speak those languages/accents natively as well as the ability to mimic regional accents and dialects (Baele & Brace, 2024). To combat this, companies have deployed matching systems that detect known content and classification systems which flag new, but similar, material (Barnes, 2022). Both approaches rely on human labour to keep up with the evolving nature of hate speech and extremist content, but the latter shows promise because of improvements in Natural Language Processing (NLP) and computer vision (Barnes, 2022).

One of the concrete examples of malign use referenced in the broader discussion on the role of AI in polarisation and political extremism is the intersection of AI-augmented data analytics with political campaigns, drawing on the Cambridge Analytica scandal and its associations with the Brexit referendum in the UK and the U.S. election in 2016. The Cambridge Analytica scandal illustrates the potential for sophisticated data mining and algorithmic micro-targeting to be used not only to tailor messaging but to intensify polarization, shape identity-based appeals, and manipulate voter segments in ways that seek to undermine democratic processes. The case is frequently cited in the across disciplinary fields to "demonstrate the power and potential of Artificial Intelligence to divide societies" if leveraged by malign actors (Burton 2023: 5).

LESSONS FROM STUDIES OF RACISM AND MISOGYNY

Studies have illustrated the ways in which word-embedding algorithms used by search engines and other platforms have racist and sexist code contained within them (Noble, 2018). Most existing computer-based facial recognition systems are trained on mugshots, photos of crime suspects, creating an effect where algorithms code these groups as inherently negative (Noble, 2018; Williams, 2024). Social media thus plays a significant role in pathways to radicalisation, with a clear and sustained increase in online extremism over time and around cross-cutting themes such as immigration, race, or LGBTQ issues (Gal et al., 2022). Compounding automated attempts to detect racism is algorithms' poor ability to detect biological racism or common cultural racism. This problem is made worse due to the high degree of normalization of culturally racist language that has relied on language that is more about exclusion, values, politics, or perceived "cultural incompatibility," most of which do not include language that violates current language guidelines on most social media platforms (Corella & Moreno, 2025). Influential users' use of coded racism sparks engagement, which algorithms reward because of the debate it often sparks (Corella & Moreno, 2025).

Research on the efficacy of algorithm-centric content moderation has shown that algorithms employed by major social media platforms are biased against people adopting distinctive communicative patterns and norms common in marginalized communities, even including entire countries (Golunova, 2025). Algorithms push users towards keywords on social media platforms that exhibit gender bias and inequality while limiting the recommendation of women and sub-cultural language used by them (Golunova, 2025). Women and other marginalized groups who create and share content expressing their perspectives on gender and other topics are subjected to algorithm-driven siloing away from users from dominant groups, effectively censoring them. Scholars also note that although major steps have been made in the development of algorithmic approaches for detecting hateful and abusive content and conversations in social media, the studies on the design of methods to detect misogynistic content about women are not as well developed. This is found to be especially true in some studies of non-English and code-mixed languages (Singh, Sharma, & Singh, 2025).

LANGUAGE POLICE ON FOR-PROFIT PLATFORMS – LESSON FROM STUDIES OF CONTENT MODERATION

On a fundamental level, the contemporary digital dynamic is that more speech on tech platforms equates to more profit for companies (Corella & Moreno, 2025). Outrage, controversy, and debate all lead to more content, and more money for social media companies. This is an unsustainable dynamic, and critical research on the topic clearly suggests that it has negative implications for healthy democracy, polarization, and extremism.

Current approaches to content moderation on major tech platforms rely on a diverse set of standards, approaches, and technologies when addressing speech. One study of Twitter (now X) and its use of Google's Perspective Application Programming Interface (API), a commonly used API, demonstrated that platform guidelines around toxic language, the algorithms it uses to detect it, and the underlying ideologies and theories that underpin the technological and social approaches to discriminatory or hateful rhetoric are ill-conceived and are undemocratic (Oh & Downey, 2024). Scholars note that algorithms and the human-created guardrails used to police toxic speech, loosely defined by various platforms as a mix of undesirable content ranging from rude, insulting, to hateful messages, are adept at identifying and removing 'uncivil' language and slurs (Oh & Downey, 2024). However, they argue that this 'uncivil language' is an essential part of democracy and undermines the ability of marginalized groups to voice their opinion when they feel they or others have been wronged (Oh & Downey, 2024).

By targeting uncivil language, existing hierarchies and systems go unchallenged with the help of tech companies. Although work is being done in the field of AI to address this, currently messages containing minority identity markers, messages with curse words, and messages with specific cultural

dialects have a higher probability to be misclassified as toxic regardless of the context in which they are expressed (Oliva, Antonialli & Gomes, 2021; Kerrigan & Barry, 2023). Instead, a focus on hate speech—speech that contributes to systemic discrimination against marginalized groups and is a key component of extremist content—is a better target for platforms (Oh & Downey, 2024). This issue of a wide, loose net approach is visible, in particular, when considering how algorithms classify and censor vernacular language used in the queer community (Kerrigan & Barry, 2023). While algorithms are not always likely to classify speech commonly used in the queer community as worthy of removal, platforms are prone to deploy the algorithmic approach of shadow banning, a set of practices designed to monitor, track and limit content that, in the subjective view of human moderators, is on the edge of breaking the platform's guidelines (Kerrigan & Barry, 2023). This not only legitimizes existing hegemonies, but also prevents representational expression and dialogue, further fuelling echo-chambers that lead to everyday extremism directed toward out-groups.

CURBING THE SPREAD OF EXTREMISM USING AI - LESSONS FROM COMPUTER SCIENCE

AI tools and changes to algorithms can help identify and moderate high volumes of extremist content that are identical, but the fast moving and nuanced of changes that can be made to this content means that people from diverse backgrounds must play an active role in the process (Golunova, 2025). Before any technical changes can be implemented effectively, social media and tech platforms must be regulated into making their "black boxes" accessible to relevant oversight parties (Williams, 2024). An increasingly small number of employees at these companies have access to the core elements of the systems that power generative AI, social media algorithms, and content moderation (Williams, 2024). This is not due to a lack of skilled employees, and like with medical research, the public might be wary of using products where the people responsible for them don't have full insight into the process.

Literature focused around curbing the spread of extremist TTI content suggests that perturbation techniques, small, deliberate changes added to imagery that is invisible or nearly invisible to the human eye and serve to confuse AI models that try and learn and copy from the content, should integrated into social media platforms as experimental features, such as optional image filters users can opt into when uploading an image (Wagner & Cetinic, 2025). Existing services that offer these perturbation techniques, such as Glaze, Nightshade, and MetaCloak, need constant updating and fine-tuning in order to perform at a high standard. Furthermore, their effectiveness is dependent on legislation and incentivizing platforms that create and share images to incorporate them (Wagner & Cetinic, 2025). Automated moderation systems used to detect visual synthetic content such as deepfakes are currently more efficient and successful at detecting, labelling, and, if the content violates platform rules, removing these forms of content than human moderators if the goal is simply to determine if what it is looking at is human or computer generated (Gandhi, 2024). However, literature is clear that the ability to label synthetic content does not stop the spread of extremist narratives or imagery, there must be stringent, cross-platform rules that place the liability on the platforms for hosting the content (Gandhi, 2024).

In-depth research about AI content detection systems actually used by major tech platforms in recent years is not currently available. However, recent literature notes that these platforms are still reliant on unimodal AI systems of content detection (Huang et al., 2025; Polli, 2025). These unimodal approaches primarily are able to detect text, leaving them especially inadequate in social media settings where images, including extremist and hateful memes made by both humans and generative AI are reliable features of mainstream discourses (Huang et al., 2025; Petruk, 2025; Polli, 2025). Content that contains both visual and verbal meanings cannot be studied by simply examining the two as separate. These multimodal messages are modified, compounded, and amplified by visual and textual combinations. In this respect, well-trained humans are far ahead of AI in determining the intended meaning.

In recent years, Multimodal Deep Learning (MDL) models have gained traction. MDLs operate by using separate neural networks to manage each type of input on its own before combining them to make

predictions or decisions. This is a resource intensive computer process, one that humans can perform nearly instantaneously, making it largely an additive tool for content moderation processes (Huang et al., 2025; Polli, 2025). Still, MDLs outperform unimodal content moderation systems, and this gap is only likely to increase in the coming years (Huang et al., 2025). Another study on computational behaviour of radicalisation shows that individuals at risk of radicalisation can be identified early by analysing both *what* they say (lexical indicators of polarisation and extremism) and *how* they engage (patterns of commenting intensity, session length, and frequency) (Gal et al., 2022). Engagement behaviour proves to be a particularly strong signal, while combining engagement and content features yields the most accurate predictions. However, technological solutions alone are insufficient. Citizen science can complement computational approaches by engaging users in real-world testing, crowdsourced monitoring, and the shared design of online spaces free from targeted extremism.

HUMAN-COMPUTER PARTNERSHIPS

A lack of shared platform guidelines and enforcement mechanisms regarding extreme content has created too many pathways for users to have their content gain traction with algorithms. Research suggests there is promise in community-centric AI-assisted moderation systems that can help curb this violent rhetoric (Udupa & Koch, 2024; Singh, Sharma, Singh, 2025). This can involve working to identify and ultimately censor or remove extremist or otherwise violent expressions using technology and carefully chosen human partners from a variety of relevant fields. By working with those engaged in the study and practice of community and technology, human-computer partnerships can identify, label, and classify contextually dependent language in concert with ethnographers, NLP researchers, and fact checkers as community partners in iterative dialogue. Although advancements in AI literacy and education are needed at all levels of society, citizen science offers a strong, proven track record mirroring this kind of partnership, with examples from AI and beyond that offer potential pathways to identify the contextual, multi-facetted ways in which extremist narratives and content impact publics (Duerinckx et al. 2024; Eliot & Bantjes, 2024; Fraisl et al. 2025). Ultimately, making the necessary changes in collaboration with the public and its people's best interests means integrating perspectives from the humanities that account for systemic discrimination, the roots of extremism, and ethics (Williams, 2025).

Working partnerships, lines of communication, and oversight between computer scientists and social scientists is critical to prevent furthering social inequalities (Kerrigan & Barry, 2023). As with any new, resource-intensive, and influential consumer product promising to address society's ills, there must be space for critical technoscience perspectives that have the regulatory capacity, and the ability to engage with the public to push for and enact changes (Kerrigan & Barry, 2023). Scholars suggest tech companies should be regulated into improving the scope, accessibility, and readability of user data. Such policies should be implemented with input from users of the site. Social media companies' algorithms should be subjected to regular audits from independent organizations and publicized. Governments and NGOs should also utilize existing consumer protections and safeguards in non-technology sectors and apply them to technology firms. Furthermore, computerized interventions into extremist narratives would be strengthened by adjusting algorithms that currently reward PRIME content. Working together with social scientists, the public, and marginalized communities, developers could integrate interventions into existing code that modulate or down-weight content with PRIME characteristics. Such technical adjustments, done in close partnership with experts in various social science fields and community members, may help reduce the unintended algorithmic amplification of extremist messaging. Like any effective content moderation, this requires strong community partnerships and constant monitoring to keep up with changes in language, technology, and shifting social dynamics. Given the tendency of extremist actors to use niche and highly specific words, phrases, and sentence structures, it is critical that experts—including citizen scientists who are able to move quickly—are empowered to identify the latest changes in extremist language trends and alert those involved in the moderation and fine-tuning of LLM and TTI models (Baele & Brace, 2024).

ALGORITHMIC INTERVENTIONS AND INTERACTIONS WITH USER BEHAVIOUR – LESSONS FROM SOCIAL PSYCHOLOGY

Algorithms play a central role in shaping online content and influencing user behaviour, particularly through their interaction with social learning processes (Brady et al.'s 2023). One of the main functions of algorithms in social media platforms is to attract and maintain user interest and engagement. To achieve this, algorithms operate in ways that promote content that both attracts and sustains user attention, interest and engagement with the platform. For these reasons, algorithms are likely to be exploiting characteristics of content that will be attracting user interest and attention. Specifically, content and information that shares the prestigious (PR), in-group (I), moral (M) and emotional (E) characteristics, or collectively PRIME characteristics, will tend to attract and maintain user attention (Brady et al., 2023). In other words, PRIME information aligns with social learning biases, and when algorithms intentionally promote such content to maximize engagement for monetization and profit-driven goals, it becomes amplified within online environments. This occurs through an interactive, iterative positive-feedback loop between users and algorithms (Brady et al., 2020). The exact workings and potential implications of this are discussed below.

Social learning theory highlights that individuals do not learn uniformly from others; rather, they prioritise information and behaviour from sources deemed prestigious or belonging to their ingroup and pay particular attention to emotional or moral content (Brady et al., 2020). These tendencies are adaptive in offline contexts, where they can promote cooperation and shared understanding. However, in online environments, algorithms exploit and reinforce these biases, often amplifying content that appeals to group identity, moral outrage, or emotional arousal. This creates feedback loops between human cognition and algorithmic optimisation, magnifying certain types of narratives—including extremist ones, at the expense of diversity and moderation. Citizen science is well-suited to help participants and the broader public visualize these loops, allowing users to trace how algorithms reward or suppress forms of extreme, computer-produced content.

Even though attentional biases towards information that has PRIME characteristics can have a positive influence on social learning by increasing effectiveness and optimal functioning in ways that enhance cooperation and improve collective problem solving, the promotion of PRIME content online may not promote the above principles. There are several reasons why this can be the case which mainly stem from the fact that PRIME characteristics are not objective but can vary depending on the contextual environment. First, prestige and status in online contexts may be linked to uncertain or dubious qualities, denote accounts with high number of followers even if of unknown identity; prestige may be linked to accounts of divisive and controversial personalities and to politicians which may be corrupt and disingenuous (see original references cited in Brady et al., 2023). Second, preference towards ingroup and negative moralised information, including political information, and its amplification in online context may be problematic as it can lose its original functions of both effectuating cooperation and eliminating non-cooperative behaviour. Amplification of ingroup content at the expense of outgroup information, can severely restrict the variety of perspectives accessed by users (similarly to eco-chamber effects, explained previously), especially when online groups in social media are formed on the basis of arbitrary social divisions, which can accentuate tunnelling of perspectives viewed online, foregoing exposure to outgroup information, with implications such as discrimination, prejudice (ingroup vs. outgroup) and the creation of a false concept of consensus. This can also lead users to establish the belief that certain perspectives disseminated online are representative of a higher proportion of the population than they are in fact in reality. Moreover, the amplification of negative moral (including political) information online, could stop serving its original purpose of preserving cooperation through, for example, the identification and resolution of behaviours that devalue established norms. Instead, negative expressions and offensive statements online can get amplified and spread even further through mutual accusations, offensive statements and comments.

EVIDENCE OF AMPLIFICATION OF PRIME CONTENT IN ONLINE CONTEXTS

There is increasing evidence that algorithms, trained upon user preferences and biases towards PRIME content, they'll also promote such content by design, leading to a concentration of attention among a small number of users. For example, Facebook and YouTube algorithms tend to promote partisan in-group content while limiting exposure to opposing viewpoints, which can further polarize users (Bisbee et al., 2022; Levy, 2021; Zhu & Lerman, 2016). Other studies identify that when out-group content is visible in such situations, it will be accompanied by in-group member commentary, which might demean and devalue such content (Brady et al., 2020) indicating an indirect way that online content gets filtered through an in-group lens. As users engage more, these algorithms intensify content recommendations, amplifying extreme views and enhancing ideological divisions.

The amplification of PRIME content in online contexts has significant implications for user behaviour, driven by both observational and reinforcement learning mechanisms. For example, social learning mainly involves learning by observation and repeating /imitating behaviour of others. Studies have shown that user expression of outrage on Twitter can be explained in parts by the amount of outrage that they have witnessed in the platform (Brady et al., 2021). Moreover, studies across platforms with regards to political information, showed that observing political content and viewpoints on a user's feed can explain their engagement with political posting online, which shows the application of mechanisms of social learning (Kim and Ellison, 2022). Kim and Ellison (2022) further argued that social learning is the key mechanism that links online and offline engagement. However, this relationship is not uniform across online users of social media. Longitudinal research shows a stronger effect over the years, a trend that is observed across countries and over time (Boulianne, 2020).

The second mechanism through which content algorithms promote PRIME content online is through processes of reinforcement learning. Algorithms reward users who post and share PRIME content with greater exposure. This implies that users may be motivated to share such content in order to enjoy social visibility and feedback. When online behaviour is rewarded then users are likely to shape future pattern of behaviour in ways that accrue greater social visibility to them (Gallego et al., 2021). Even though the vast majority of users are unaware of the impact of such algorithmic interaction with user content in shaping subsequent online behaviour, a few occasions have been observed where users alter their online activity intentionally in ways that help them to benefit from the algorithmic impact. For instance, individuals that use online accounts for personal promotion for commercial or political reasons, can capitalise on this by manipulating their content to feature more prominently on user feeds (this is also known as 'gaming'). This in turn can lead to certain users gaining outsized influence on social media and to niche political views being perceived as having greater visibility and reach than in reality. For instance, a study of a subreddit (r/the_donald) based on 2016 data shows how strategic manipulation of Reddit's sorting algorithm took advantage of algorithmic nuances and used the sorting process rhetorically to alter a dominant argument (Shepherd, 2020). This leads to argument amplification on the platform but also increases the chances of arguments being carried over to other platforms. Hence, the tendency of algorithmic-user interaction to lead to spreading of outrage and anger online can be capitalised by individuals who may intentionally promote content that steers negative emotions online, in order to gain higher visibility of certain political views.

RECOMMENDATIONS FOR PUBLIC ENGAGEMENT

Engaging the public in addressing the issues raised within this handbook requires an inclusive and proactive approach. We propose a citizen science model as a means to empower individuals in navigating AI and the technological and social-psychological dynamics of social media. By enhancing users' understanding of how algorithms shape their online interactions and influence their perceptions, we can foster a more critical and self-aware public that actively engages with digital platforms in a way that mitigates the risks of radicalization and misinformation. The challenge, then, lies in recognising how subtle, often normalized patterns of discourse—such as humour, grievance, and distrust—can serve as conduits for violence and systemic discrimination, shifting the boundaries of what counts as acceptable speech and belief in liberal democracies. By fostering public awareness and participation and by equipping citizens with the knowledge and tools they need to critically assess such content, we can begin to understand and dismantle these harmful patterns before they escalate into violence.

This final section of the handbook outlines how citizen science and crowd-based research can serve as an open inclusive tool for increasing public awareness and engagement with complex socio-technical issues, such as disinformation, algorithmic bias, and the circulation of extremist narratives online. By involving participants directly in the processes of data collection, analysis, and reflection, such initiatives move beyond traditional top-down models of information dissemination and instead foster critical media literacy and participatory inquiry.

Drawing on insights from our work and existing research, we focus on three core aspects in the following: Citizen-science and crowd-based methods, greater awareness of algorithmic interventions and attentional biases, and the development of AI literacy.

CITIZEN SCIENCE APPROACHES AND CROWD BASED RESEARCH TOOLS

Citizen science are collaborative approaches to research that involves non-professional participants, often the general public, in scientific investigations. These individuals and their communities contribute to data collection, analysis, and problem-solving, working alongside researchers to address specific problems and societally relevant issues. Citizen science empowers people to actively participate in scientific discovery, broadening the scope of research and increasing public engagement with science.

Citizen science has the potential to increase public awareness particularly around issues that are intangible and difficult to communicate. Actively engaging with concepts like AI, user behaviour, narratives and discourse, or platform dynamics with expert support and participatory exploration supports participants developing awareness and digital literacy.

A key component of citizen science projects is their capacity to challenge or illuminate issues that are often dominated by experts, particularly in cases where credentialed actors omit or downplay health concerns raised by communities affected by pollution (Eliot and Bantjes, 2024). One study found that a Belgian citizen science programme, amai!, which aims to democratise artificial intelligence, produced encouraging results. However, there remain few examples of such participatory approaches within the specific subject area addressed by this report (Duerinckx et al., 2024).

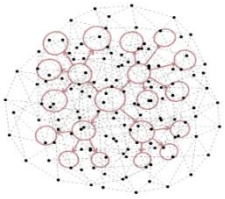
This calls for new projects that adopt a stronger and more critical approach to the role of artificial intelligence in the pursuit of just societies, a perspective that until recently has been largely absent from expert discourse. Universities and their researchers, supported by both public and private funding, have played a significant role in shaping the development, narratives, and adoption of artificial intelligence within the private sector and the public sphere (McCorduck, 2004). In light of the issues outlined in this handbook concerning the ways in which artificial intelligence and algorithms contribute to the spread and monetisation of extremism, it is essential that the same experts and institutions who helped bring about the current technological landscape are not the sole voices involved in identifying solutions intended to serve the public interest

Citizen science pursues a range of simultaneous strategies to tackle this. Firstly, citizen science projects engage users in shaping and determining the research questions of specific projects, so that they involve topic areas that are better tuned in addressing emerging problems. This is particularly crucial for topic areas and research fields that are currently less accurately defined or that are still evolving and requiring greater and more diverse inputs to understand their nature and depth. Furthermore, citizen science projects can leverage participants to offer resources, such as their time and expertise or to get access to diverse material and data. These contributions can lead to establishing unique datasets more quickly and effectively as compared to 'closed' projects. For instance, the 'YouTube Regrets' project by Mozilla is a crowdsourced browser extension with over 37 thousand volunteers which document and report recommendations of videos that they 'regret' watching on the platform, or otherwise video recommendations that they consider harmful (Matias, 2023; Mozilla, 2021).

Additionally, these projects could potentially lead to a change in user knowledge, skills and attitudes and potentially shape future online behaviour. Ultimately this can facilitate broader changes in socio-technical systems, that offer the framework for the development and application of technologies with explicit consideration of their social impact. Changes in attitudes can alter user online behaviour (being aware of attentional biases can lead to change in perception, behaviour and actions), increase advocacy for greater transparency in algorithmic use and for tighter regulatory frameworks that will be enforced by regulators and be implemented by businesses and platform owners / operators.

The case of the *Young Citizen Scientists Against Disinformation* project exemplifies this approach by engaging young adults in examining how disinformation operates on social media and in co-developing research questions that reflect their lived experiences. It was launched in July 2024 by the University for Continuing Education Krems and the FH St. Pölten and is a participatory initiative aiming at

engaging young adults in examining the dynamics of disinformation on social media (Pospisil, 2024). Through collaboration with researchers, participants help define research questions based on their own experiences and perspectives, fostering a bottom-up approach to the study of misinformation. The project encourages young adults to actively collect, analyse, and evaluate (dis)information in their everyday lives, while also providing educational resources to enhance their critical thinking and research skills. By facilitating discussions around disinformation, regulatory responses, and digital tools, the project cultivates a deeper understanding of the issue and empower participants to develop strategies for detecting and assessing the truthfulness of online content. The findings and discussions contributed to the creation of a platform offering advice, support, and practical tools for identifying and addressing disinformation, while also identifying unmet needs and concerns among young people in this area.



IMPROVING AWARENESS OF ALGORITHMIC INTERVENTIONS AND USER ATTENTIONAL BIASES

The example of how citizen participation in research projects can increase awareness on disinformation is highly relevant to organisations and NGOs seeking to develop projects to engage broader publics in neighbouring debates around emerging technologies and the impact of AI on our democratic conversations. Projects can be developed to inform and support citizens in identifying the ways in which extremist narratives circulate and gain visibility in social media and the ways algorithmic interventions and knowledge attention systems work on social media. This is pertinent to awareness about the role of algorithms in shaping online content as well as the role of user behaviour on social media and how it interacts with algorithmic ranking and recommendation dynamics.

Using case studies of how extremist narratives propagated online can demonstrate how through the interaction of algorithmic interventions and user attentional biases, online content becomes increasingly selective in ways that do not necessarily represent majority views. In this way, users-participants can be empowered to identify similar niche and extreme narratives in future situations and become more aware of how their individual choices and preferences in social media shape or distort online experiences. Users can be versed on how to use filtering tools or other such tools that can increasingly become more widespread and available on social media platforms. The issue of trust in information is very important, and the aim is to increase online user critical approach towards social media stories and narratives, encouraging them to cross-check information across platforms or to use established news outlets.

An example of such an approach is the data donation project DataSkop, which investigates platforms for democratic engagement, with a particular focus on the German federal election of 2021. The project began as a pilot study examining the personalised recommendations and search results delivered by YouTube in relation to election related topics. In 2023, a similar initiative focused on TikTok's recommender system. Across both projects, volunteer contributed data were used to produce reports analysing how these recommender systems operate (DataSkop, 2025).

Moreover, such datasets could in turn provide the inputs for future projects through open access and sharing protocols. For instance, data providing examples of disinformation can be used to train algorithms that will automatically, or with some involvement from moderators, identify and highlight content that after moderation by experts could almost automatically be eliminated from platforms. This process is already in place on some platforms, but the development and further refinement of relevant algorithms can be facilitated by broader and more diverse training data, which can be obtained through CS projects.

Gradually users-participants can become aware of how algorithmic sorting primarily serves profiteering interests and secondary purposes of 'personalisation' of content. Moreover, users can

become aware of own attentional biases towards PRIME content and particularly in terms of evaluating concepts such as 'prestige'. Awareness around attending to negative and emotive content is not widespread across the online user community, with the use of participant engagement projects potentially enabling participants to develop such awareness, changing over time their online behaviour, reactions and responses towards online content.

A key step forward is to promote greater transparency in how algorithms are designed and operate, including clearer disclosure of the training data that underpin them. Equally important is fostering public understanding of how individual attentional biases interact with algorithmic sorting processes to shape online experiences. Raising such awareness can help users recognise their own role in reinforcing engagement-driven systems. In parallel, platforms could implement interventions that give users more meaningful control over what they see, such as personalised filtering tools that allow individuals to curate their news feeds intentionally rather than passively consuming algorithmic selections. The following sections illustrate how this can be advanced in practice through approaches drawn from open science and citizen science initiatives.

BUILDING AI LITERACY

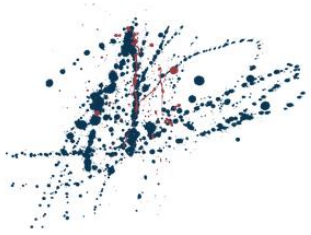
AI literacy has become an essential skill for navigating the complexities of social media platforms and the broader digital ecosystems of which they are part. As algorithms play an ever-larger role in shaping what users encounter online, understanding how these systems work and the biases they may introduce is critical. AI literacy is not just about knowing how algorithms and the mechanisms behind AI operate technically but also about recognizing their social and psychological implications. This knowledge empowers individuals to critically engage with digital content, making them less susceptible to manipulation, echo chambers, or extremist ideologies.

One key aspect of AI literacy is understanding attentional biases and how they interact with algorithmic sorting processes. Algorithms are designed to prioritise content that grabs attention, often amplifying emotionally charged or sensational material. This dynamic can reinforce existing beliefs and fuel polarization. When users are aware of how these mechanisms work, they can more consciously curate their digital experiences, helping them break free from filter bubbles and identify potentially harmful content before it influences their views. Furthermore, fostering this awareness encourages users to take a more active role in shaping their own digital environments, leading to healthier online spaces.

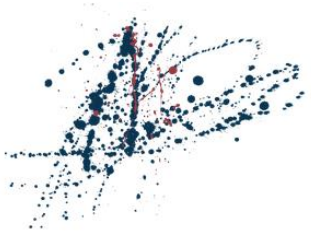
To build AI literacy, education on algorithms and digital media literacy has to be integrated into school curricula, public workshops, and digital platforms. By offering users the tools to recognize the invisible forces shaping their online interactions, we can foster a more informed and resilient digital citizenry that helps people navigate the challenges of an algorithm-driven world with greater critical awareness and responsibility.

KEY TAKE-AWAYS

For practitioners designing public engagement initiatives around AI and extremism, the discussions in this handbook offer five key insights:

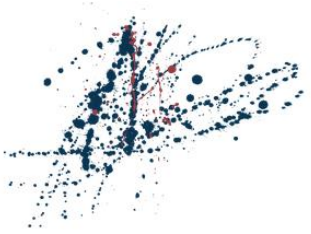


As with disinformation, tackling extremist content requires not only debunking false claims but also addressing the **algorithmic and attentional dynamics** that amplify certain narratives. Drawing from crowd science, practitioners can design projects that make these mechanisms visible illustrating how recommendation systems and user engagement biases contribute to the disproportionate prominence of extreme views.



Promote **AI literacy** and **critical thinking skills**: Educating the public on how attentional biases influence their interactions with algorithms is crucial. By fostering media literacy and critical thinking, users can better recognize the role these biases play in shaping their online experiences, helping them make more conscious, informed decisions about the content they engage with.

Social media platforms and tech-firms need to be held accountable and develop transparent and ethical algorithm design. Policy makers and the general public should advocate for the development of more **transparent and ethically designed algorithms** that minimize harmful biases and prioritize diverse, balanced content. By involving the public in discussions around algorithmic accountability, users can be empowered to demand more responsible design that reduces the potential for reinforcing echo chambers or extremist ideologies.



The **participatory and educational dimensions** of citizen science parallel successful disinformation interventions: they foster critical literacy, empower individuals to question content sources, and encourage cross-platform verification. This mirrors the shift in disinformation research from fact-checking to **building epistemic resilience** - the ability of individuals and communities to critically engage with and evaluate digital information ecologies.

The handbook underscores the broader **societal benefits of participatory research,** such as generating new datasets, promoting digital accountability, and supporting behavioural change. These mechanisms are directly transferable to initiatives seeking to counter extremist narratives by turning users into co-investigators of the social and algorithmic conditions that enable extremism to thrive. In essence, crowd-based awareness projects can transform passive consumers of online content into **active, critically aware participants** in shaping ethical and transparent AI-mediated public discourse.

LITERATURE

Akerlund, M. (2021) Dog whistling far-right code words: The case of culture enricher' on the Swedish web, *Information, Communication & Society* 25(12), 1808-1825, https://doi.org/10.1080/1369118X.2021.1889639

Aljabri, M. et al. (2023). Machine learning-based social media bot detection: a comprehensive literature review. *Social Network Analysis and Mining,* 13(20), https://doi: 10.1007/s13278-022-01020-5

Allchorn, W. (2024). Global Far-Right Extremist Exploitation of Artificial Intelligence and Alt-Tech. *Counter Terrorist Trends and Analyses*, *16*(3), pp. 13-18. https://www.jstor.org/stable/48778662.

Andreouli, E. et al., (2024). Deliverable D5.1. Framework paper on emergence of opposition drivers across sites and shared dialogical interventions. Available at: https://oppattune.eu/wp-content/uploads/2025/06/D5.1.pdf

Askanius, T., Haselbacher, M., Reeger, U., and Stoencheva, J. (2024a). Deliverable D4.1. Visualisation report of emerging extremist narratives across Europe. Available at: https://oppattune.eu/wp-content/uploads/2024/05/D4.1_OppAttune-_FINAL.pdf

Askanius, T., Haselbacher, M., Reeger, U., and Stoencheva, J. (2024b). Deliverable D4.2. Translocal articulations of extremism in three countries. Visualisation report on Austria, Bulgaria, and Sweden. Available at: https://oppattune.eu/wp-content/uploads/2025/06/D4.2.pdf

Awasthi, S. (2025) *From clicks to chaos: How social media algorithms amplify extremism*, *orfonline.org*. Available at: https://www.orfonline.org/expert-speak/from-clicks-to-chaos-how-social-media-algorithms-amplify-extremism.

Baele, S. and Brace, L. (2024) *AI Extremism: Technology, Tactics, Actors*, VOX-Pol. Available at: https://voxpol.eu/wp-content/uploads/2024/04/DCUPN0254-Vox-Pol-AI-Extremism-WEB-240424.pdf.

Baele, S.J., Naserian, E. and Katz, G. (2024). Is AI-Generated Extremism Credible? Experimental Evidence from an Expert Survey. *Terrorism and Political Violence*, 37(8), pp. 1060-1076, https://doi.org/10.1080/09546553.2024.2380089.

Bandara, C. (2024). Hallucination as disinformation: The role of LLMs in amplifying conspiracy theories and fake news. *Journal of Applied Cybersecurity Analytics, Intelligence, and Decision-Making Systems*, 14(12), pp. 65-76, https://sciencespress.com/index.php/JACAIDMS/article/view/14.

Barnes, M. R. (2022) "Online Extremism, AI, and (Human) Content Moderation", *Feminist Philosophy Quarterly*, 8(3/4), http://doi.org/10.5206/fpq/2022.3/4.14295.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim Code*. Polity Press

Berger, J.M. (2018). *Extremism*. Cambridge: MIT Press.

Bisbee, J., Brown, M., Lai, A., Bonneau, R., Tucker, J. A. & Nagler, J. (2022). Election fraud, YouTube, and public perception of the legitimacy of President Biden. *Journal of Online Trust and Safety,* 1(3), https://doi.org/10.54501/jots.v1i3.60

Boulianne, S. 2020. Twenty Years of Digital Media Effects on Civic and Political Participation. *Communication research,* 47, 947–966, https://doi:10.1177/0093650218808186.

Brady, W. J., Crockett, M. J. & Van Bavel, J. J. 2020. The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online. *Perspectives on Psychological Science,* 15, 978–1010, https://doi:10.1177/1745691620917336.

Brady, W. J., Jackson, J. C., Lindström, B. & Crockett, M. J. 2023. Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences,* 27, 947–960, https://doi:10.1016/j.tics.2023.06.008.

Brady, W. J., Mcloughlin, K., Doan, T. N. & Crockett, M. J. 2021. How social learning amplifies moral outrage expression in online social networks. *Science Advances,* 7, eabe5641, https://doi:10.1126/sciadv.abe5641.

Burton, J. (2023) Algorithmic extremism? The securitization of artificial intelligence (AI) and its impact on radicalism, polarization and political violence, *Technology in Society,* 75, 102262, https://doi.org/10.1016/j.techsoc.2023.102262.

Cleveland-Stout, N. (2025) *Israel wants to train ChatGPT to be more pro-Israel*. Responsible Statecraft, 29 September. Available at: https://responsiblestatecraft.org/israel-chatgpt/ (Accessed: 2 October 2025).

Conway, M. (2017). Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research, *Studies in Conflict & Terrorism*, 40(1), pp.77–98.

Corella, Á.S. and Moreno, N.H. (2025). Racism in the Digital Age: The Impact of Social Media Algorithms on Public Discourse. *The Age of Human Rights Journal*, *25*, e9603. https://doi.org/10.17561/tahrj.v25.9603.

Dataskop. 2025. *What is DataSkop?* [Online]. Available: https://dataskop.net/overview-in-english/ [Accessed 12 August 2025].

Delkhosh, F. et al. (2023). 'Impact of Bot Involvement in an Incentivized Blockchain-Based Online Social Media Platform', *Journal of Management Information Systems*, 40(3), pp. 778–806, https://doi.org/10.1080/07421222.2023.2229124.

Duerinckx, A. et al. (2024) 'Co-creating Artificial Intelligence: Designing and Enhancing Democratic AI Solutions Through Citizen Science', *Citizen Science: Theory & Practice*, 9(1), p. 43, https://doi.org/10.5334/cstp.732.

Ebner, J. (2023) *Going mainstream: how extremists are taking over*. London: Ithaka.

Eliot, D. and Bantjes, R. (2024). "Climate science vs denial machines: how AI could manufacture scientific authority for far-right disinformation" in I. Allen *et al.* (eds.) *Political Ecologies of the Far Right: Fanning the Flames*. Manchester, UK: Manchester Press (Global Studies of the Far Right), pp. 213-231.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.

Fisher, S.A., Howard, J.W. and Kira, B. (2024). Moderating Synthetic Content: The Challenge of Generative AI, *Philosophy & Technology*, 37(4), 133, https://doi.org/10.1007/s13347-024-00818-9.

Fraisl, D. *et al.* (2025) "Leveraging the collaborative power of AI and citizen science for sustainable development," *Nature Sustainability*, 8(2), pp. 125–132, https://doi.org/10.1038/s41893-024-01489-2.

Gallego, A., Schöll, N. & Le Mens, G. 2021. Politician-citizen interactions and dynamic representation: Evidence from Twitter. *Economics Working Papers 1769,* Department of Economics and Business, Universitat Pompeu Fabra.

Gal, K., Ravid, E., Solomon, S. (2022). Deliverable D6.1. Computational Behaviour of Radicalisation. D.Rad Project. Available at: https://cordis.europa.eu/project/id/959198/results

Gandhi, M. (2024) *Terrorism, Extremism, Disinformation and Artificial Intelligence: A Primer for Policy Practitioners*, Institute for Strategic Dialogue. Available at: https://www.isdglobal.org/wp-content/uploads/2024/01/Terrorism-extremism-disinformation-and-artificial-intelligence_A-primer-for-policy-practitioners.pdf.

Goetz, J., Sedlacek, J., and Winkler, A. (eds.) (2018). *Untergangster des Abendlandes: Ideologie und Rezeption der rechtsextremen 'Identitären'*. 2nd edition. Hamburg: Marta Press.

Golunova, V. (2025). A Misogynistic Glitch? A Feminist Critique of Algorithmic Content Moderation. *Law, Technology and Humans* 7(2), pp. 40-52, https://doi.org/10.5204/lthj.3848.

Górska, A.M. and Brzeziński, D. (2025) "AI Racial Bias: How Text-to-Image Artificial Intelligence Generators Construct Prestigious Professions," in D. Brzeziński *et al.* (eds.) *Algorithms, Artificial Intelligence and Beyond: Theorising Society and Culture of the 21st Century*. 1st ed. Oxon, UK: Routledge (Antinomies), pp. 211–226.

Handfield, T. (2023). Regulating Social Media as a Public Good: Limiting Epistemic Segregation, *Social Epistemology*, 38(6), pp. 743–758, https://doi.org/10.1080/02691728.2022.2156825.

Harris, M. (2025) *What's Left: Three Paths Through the Planetary Crisis*. New York: Little, Brown and Company.

Huang, L. *et al.* (2025). RU-AI: A Large Multimodal Dataset for Machine-Generated Content Detection, *Companion Proceedings of the ACM on Web Conference 2025*. ACM, pp. 733–736. https://doi.org/10.1145/3701716.3715306.

Kalai, A.T. *et al*. (2025) Why language models hallucinate. *arXiv* preprint arXiv:2509.04664. https://doi.org/10.48550/arXiv.2509.04664.

Kerrigan, P. and Barry, M. (2023) "Automating vulnerability: Algorithms, artificial intelligence and machine learning for gender and sexual minorities." In P. Aggleton *et al* (eds.) *Routledge Handbook of Sexuality, Gender, Health and Rights.* 2nd ed. Oxon, UK: Routledge, pp. 164-174.

Kim, D. H. & Ellison, N. B. (2022). From observation on social media to offline political participation: The social media affordances approach. *New Media & Society,* 24, 2614–2634, 10.1177/1461444821998346.

Levy, R. E. (2021). Social Media, News Consumption, and Polarization: Evidence from a Field Experiment. *American Economic Review,* 111**,** 831–70, https://doi.10.1257/aer.20191777.

Lewis, R. (2018). *Alternative Influence: Broadcasting the Reactionary Right on YouTube.* New York: Data & Society.

Lopes, H. (2024) *Melodies of Malice: Understanding How AI Fuels the Creation and Spread of Extremist Music*. GNET, 11 December [online]. Available at: https://gnet-research.org/2024/12/11/melodies-of-malice-understanding-how-ai-fuels-the-creation-and-spread-of-extremist-music/.

Matias, N. J. 2023. Humans and algorithms work together — so study them together. *Nature,* 617**,** 248–2023, https://doi.org/10.1038/d41586-023-01521-z.

Mozilla. 2021. *YouTube Regrests: A crowdsourced investigation into YouTube's recommendation algorithm* [Online]. Available: https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf [Accessed 11 August 2025].

Maynor, J. (2025) Flood the Zone with Shit: Algorithmic Domination in the Modern Republic, *Social Sciences*, 14(6), https://doi.org/10.3390/socsci14060391.

McCorduck, P. (2004) *Machines who think: A personal inquiry into the history and Prospects of Artificial Intelligence*. 2nd edn. Boca Raton: CRC Press, Taylor & Francis Group.

Midttun, J. S. (2023) 'Democratizing Social Media', *Platform Cooperativism Consortium* [blog], 21 December. Available at: https://platform.coop/blog/democratizing-social-media/

Munk, T. (2024). *Far-Right Extremism Online: Beyond the Fringe* (1st ed.). Routledge. https://doi.org/10.4324/9781003297888

Noble, S. U. (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, USA: New York University Press.

Oh, D. and Downey, J. (2024) 'Does algorithmic content moderation promote democratic discourse? Radical democratic critique of toxic language AI', *Information, Communication & Society*, 28(7), pp. 1157–1176, Available at doi: 10.1080/1369118X.2024.2346531.

Oliva T.D., Antonialli. D.M. and Gomes, A. (2021) "Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online," *Sexuality & Culture*, 25(2), pp. 700–732, https://doi.org/10.1007/s12119-020-09790-w.

Osler, L. (2025) Hallucinating with AI: AI Psychosis as Distributed Delusions. *arXiv* preprint arXiv:2508.19588, https://doi.org/10.48550/arXiv.2508.19588

Petruk, A. (2025). Memes vs. Machines: Comparison of AI-generated Images vs. Traditional Memes in Right-Wing Social Media Discourse. *Crossings: An Undergraduate Arts Journal*, 5(1), pp. 99-114, https://doi.org/10.29173/crossings324

Polli, C. (2025). A multimodal critical perspective on the challenges of automatic detection of hate speech in Internet memes: The case of simianization. *Critical Approaches to Discourse Analysis Across Disciplines*, 17(1), pp. 96-117, https://doi.org/10.21827/cadaad.17.1.42413

Pospisil, B. 2024. *Young citizen scientists against disinformation* [Online]. Available: https://www.citizen-science.at/en/projects/young-citizen-scientists-against-disinformation [Accessed 22 July 2025].

RAN (Radicalisation Awareness Network). (2019). Preventing radicalization to terrorism and violent extremism: Delivering coutner- or alternative narratives.

Rottmann et al. (2025). Deliverable D5.2. Five Country Good Practice Case Studies Report in facilitating social dialogue for resilient democracies, including depicting successful social dialogue practices that may dissolve tendencies of polarising and extremism. Available at: https://oppattune.eu/wp-content/uploads/2025/06/D5.2.pdf

Shepherd, R. P. 2020. Gaming Reddit's Algorithm: r/the_donald, Amplification, and the Rhetoric of Sorting. *Computers and Composition,* 56, 102572, https://doi.org/10.1016/j.compcom.2020.102572.

Shin, D. (2024). "Misinformation, Extremism, and Conspiracies: Amplification and Polarization by Algorithms," in *Artificial Misinformation: Exploring Human-Algorithm Interaction Online*. Cham: Springer Nature Switzerland, pp. 49–78, https://doi.org/10.1007/978-3-031-52569-8_3

Shin, D. and Jitkajornwanich, K. (2024). How Algorithms Promote Self-Radicalization: Audit of TikTok's Algorithm Using a Reverse Engineering Method, *Social Science Computer Review*, 42(4), pp. 1020–1040, https://doi.org/10.1177/08944393231225547.

Singh, A., Sharma, D. and Singh, V.K. (2025). Misogynistic attitude detection in YouTube comments and replies: A high-quality dataset and algorithmic models. *Computer Speech & Language*, *89*, https://doi.org/10.1016/j.csl.2024.101682.

Solomon, S., Spanò, G. (2023). Deliverable D6.5. Artificial Intelligence, Human Rights and Civil Liberties. D.Rad Project. Available at: https://cordis.europa.eu/project/id/959198/results

Stoencheva, J., Haselbacher, M., Askanius, T. and Reeger, U. (2026) Everyday extremism: Memes and the mainstreaming of extremist narratives around the 2024 European Parliament Election, *Journal of Popular Communication*, forthcoming spring 2026

Tencer, D. (2024). *$125m-backed Suno is being used to make racist and antisemitic music*. Music Business Worldwide, 20 June [online]. Available at: https://www.musicbusinessworldwide.com/125m-backed-suno-is-being-used-to-make-racist-and-antisemitic-music/.

Tondo, L. (2025). *Italian opposition file complaint over far-right party's use of 'racist' ai images*, *The Guardian*. Available at: https://www.theguardian.com/technology/2025/apr/18/italian-opposition-complaint-far-right-matteo-salvini-lega-racist-ai-images

Udupa, S. and Koch, L. (2024). Tackling online misogyny in political campaigns: promise and limitations of artificial intelligence', *Feminist Media Studies*, 24(6), pp. 1428–1434. https://doi.org/10.1080/14680777.2023.2291313

Van Der Ven, H. et al. (2024). Does artificial intelligence bias perceptions of environmental challenges? *Environmental Research Letters, 20*(1), 014009. https://doi.org/10.1088/1748-9326/ad95a2.

Verma, P. (2024). "These ISIS news anchors are AI fakes. Their propaganda is real" *The Washington Post*, 17 May. Available at: https://www.washingtonpost.com/technology/2024/05/17/ai-isis-propaganda/.

Wagner, L. and Cetinic, E. (2025). Perpetuating Misogyny with Generative AI: How Model Personalization Normalizes Gendered Harm. *arXiv preprint arXiv:2505.04600*. https://doi.org/10.48550/arXiv.2505.04600

Williams, D.P. (2024) 'Disabling AI: Biases and values embedded in artificial intelligence', in D.J. Gunkel (ed.) *Handbook on the Ethics of Artificial Intelligence*. Cheltenham, UK: Edward Elgar Publishing, pp. 246–261.

Wyer, S. and Black, S. (2025) Algorithmic bias: sexualized violence against women in GPT-3 models, *AI and Ethics*, 5(3), pp. 3293–3310. https://doi.org/10.1007/s43681-024-00641-0.

Zhang, X., & Davis, M. (2024). E-extremism: A conceptual framework for studying the online far right. *New Media & Society*, 26(5), 2954–2970. https://doi.org/10.1177/14614448221098360

Zhu, L. & Lerman, K. 2016. Attention inequality in social media. *arXiv preprint arXiv:1601.07200*, https://doi.org/10.48550/arXiv.1601.07200