SITRA

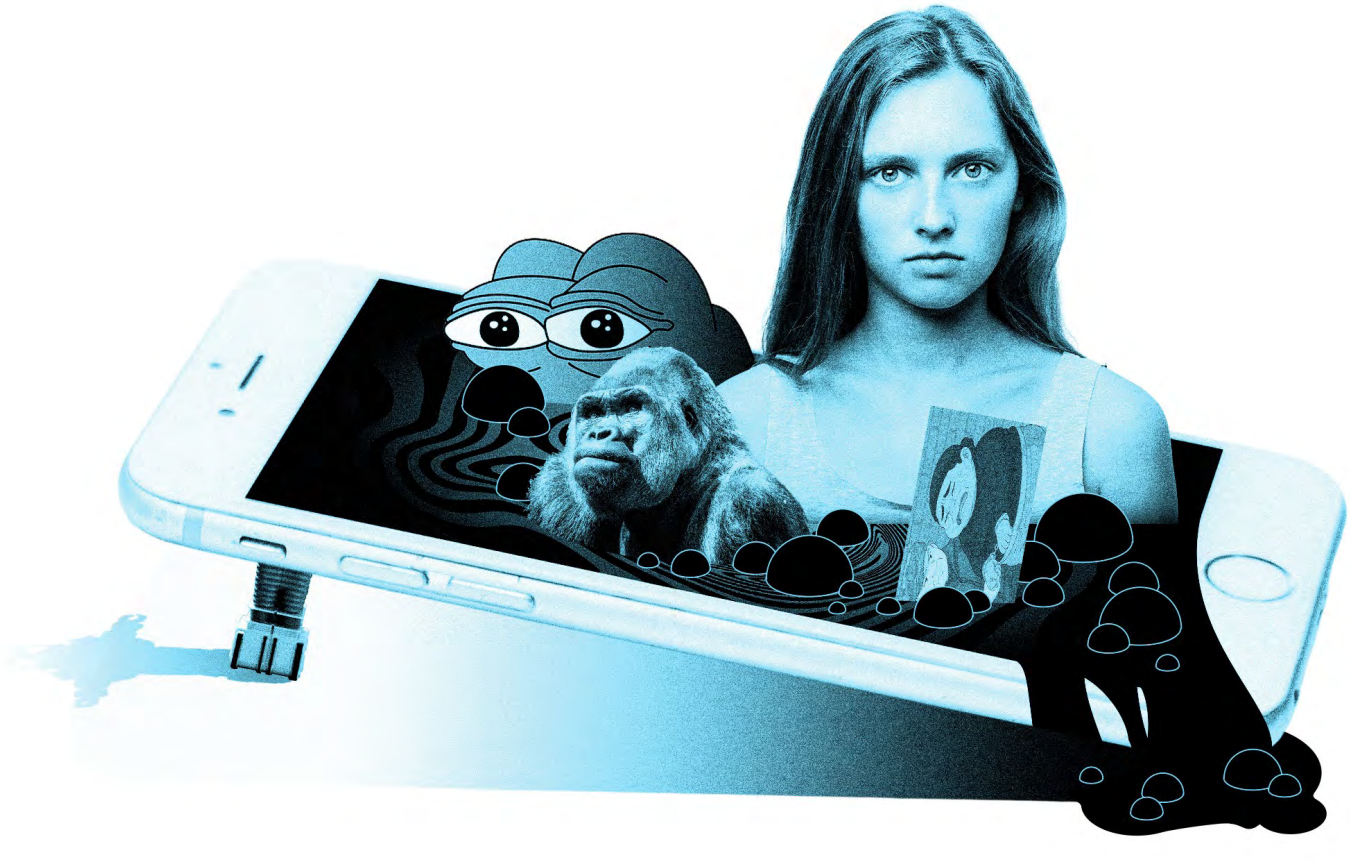
SITRA STUDIES 256

MARCH 2026

ALGORITHMS AND DEMOCRACY

How social media shapes young Europeans' worldviews

Ilkka Räsänen, Kristo Lehtonen, Bálint Dercsényi, Laurence Fenn, Sujatha Krishnan-Barman & Cindia Li



Sitra studies 256

Algorithms and Democracy
How social media shapes young Europeans' worldviews

Working group at Sitra: Pipsa Havula, Tiina Härkönen,
Harri Junttila, Kristo Lehtonen, Ilkka Räsänen, Jukka Vahti

Cover image: Topias Dean

Layout: Grano Oy

ISBN 978-952-347-457-4 (PDF) www.sitra.fi
ISSN: 1796-7112 (Sitra Studies) (online publication) www.sitra.fi

THE SITRA STUDIES publication series features the results of Sitra's future-oriented work and experiments.

Contents

Foreword

Social media platforms have rapidly become the central gateways of the digital age. These platforms influence behaviour and steer civic discourse through opaque algorithmic mechanisms. For young people, social media is now the primary source of news and political information. The content they encounter is optimised for engagement rather than democratic discourse. Platforms act as private governors of public conversations by setting their own rules and developing algorithms to determine which voices are amplified or suppressed. In this system, users are not customers or citizens but products; their data fuels advertising-driven business models that reward divisive and emotionally charged content over dialogue and compromise.

Sitra's Megatrends 2026 report highlights how democracy is being tested by authoritarian actors and ideas, declining trust, a fragmented information environment, and increasing polarisation. At the same time, rapid technological change, including generative AI, is reshaping the foundations of society, from the production of information to how influence is exercised. In this landscape, the question is no longer whether digital platforms affect democracy, but how to ensure that they support healthy democracy, rather than contribute to democratic backsliding.

Sitra's strategic objective is to protect and renew European democracy. Through our impact work on democracy and trust, we aim to make social media safer for democracy by combatting political disinformation and interference in civic discourse and electoral processes.

This publication continues Sitra's work on understanding digital power, notably Tracking the Digipower study of 2022, which examined how data can be used to predict and influence behaviour. Our purpose is straightforward: to shed light on how social media shapes young people's exposure to political information and public debate, and to help platforms, regulators, educators, families, and society take informed action to make the digital public sphere safer for democracy.

We would like to extend our warmest thanks to The Behavioural Insights Team, Bondata, Paula Gori, and the many experts who contributed their time and insights to this work.

18 February 2026

Kristo Lehtonen and Ilkka Räsänen
Kristo Lehtonen is Director of International Programmes,
and Ilkka Räsänen is Head of EU Affairs at Sitra

Commentary: Safeguarding citizens' agency

Social media platforms have evolved into public arenas where citizens engage in dialogue and share information. These digital spaces have fundamentally transformed public discourse, the range of participants involved, the format of content, and the ways in which information is accessed and consumed. Additionally, this transformation has sparked a lively debate about the implications of such spaces being owned by profit-driven organisations and the broader issue of European digital sovereignty.

The freedom to hold opinions and to receive and disseminate information and ideas is a cornerstone of democracy. Yet, for public civic spaces to function effectively, they must be founded on transparency and accountability, with the integrity of information assured. Regrettably, this standard is not what is happening on the major social media platforms at present.

This Sitra study and the platform audit conducted by the Behavioural Insights Team offers valuable insight into the nature of political content encountered by young Europeans on leading social media platforms. Political content falls under the EU Digital Services Act (DSA) risk category of civic discourse, which is central to democratic life. When participating in civic discourse within these public arenas, we are not merely users but citizens. It is essential that we retain agency over the content we access and ensure that our debates take place in environments that are safe, equitable, transparent, and accountable, and are rooted in democratic principles.

This study is particularly noteworthy for several reasons. Firstly, it substantiates previous research indicating that right-wing political content is disproportionately promoted on social media. The political material viewed by the avatars in the study did not include paid political advertising, raising the question of why content from one end of the political spectrum is recommended more frequently than the other, regardless of the avatars' online behaviour. These findings are consistent with other research suggesting that content which is emotionally charged or provocative tends to go viral more readily and consequently holds greater economic value for the platforms.

Another significant finding from the study is the unpredictable nature of algorithmic behaviour. The avatars' engagement signals, or online actions, did not have a consistent effect on the recommendations they received. Put simply, users cannot reliably anticipate or influence which content the algorithm will choose to present to them. This unpredictability is particularly concerning given that, according to the study, there were instances where avatars were exposed to extreme content following abrupt shifts, without any obvious trigger from their activity.

The direct consequence of this is that citizens' agency is undermined. When citizens' agency in choosing how they want to stay informed is undermined, both information integrity and the democratic process are compromised. Citizens are exposed to content they have neither requested nor expressed interest in, undermining both their autonomy and the diversity of information available.

Paula Gori
Secretary General and Coordinator, European Digital Media Observatory (EDMO)
European University Institute

Summary

The Finnish Innovation Fund Sitra commissioned The Behavioural Insights Team (BIT) and Bondata to examine the political content encountered by 18–24-year-olds on social media and assess its potential risks to civic discourse. Such risks are among the systemic risks identified in the European Union's Digital Services Act.

BIT conducted a systematic audit using 24 avatars on Instagram, TikTok and X in Finland, France and Romania. Avatars simulated different levels of engagement in political content, including a 'Tilted trajectory' phase where they signalled exclusive interest in either left-of-centre or right-of-centre content. In total, 1,719 political posts were manually coded for political leaning and problematic content, including misinformation, conspiracy theories and hate speech.

Bondata complemented the audit with an online survey of 3,063 young adults in the same countries, examining social media use and exposure to problematic political content.

Right-wing content dominated in the BIT platform audit, accounting for 58 per cent of all politically classified posts, compared with 26 per cent left-wing and 16 per cent centrist content. This dominance often persisted even during the 'Tilted trajectory' phase when avatars signalled interest in left-wing politics, suggesting a disproportionate amplification of right-wing perspectives regardless of user preferences. Romanian feeds were an exception: they were largely dominated by centrist content, particularly government communications. Bondata's survey complemented these findings and found that 44 per cent of users in Finland who strongly identify with the left felt that the content they received corresponded very poorly to their views, compared with only 5 per cent of those strongly identifying with the right, with similar results in France and Romania.

The audit revealed algorithmic unpredictability and a lack of user control. Engagement signals had no consistent or reliable impact on the content delivered, and feeds could shift suddenly and substantially without any clear trigger.

The majority of problematic content observed did not violate community guidelines. Outright misinformation, conspiracy theories, and hate speech were relatively rare in the BIT platform audit. Instead, feeds were dominated by unverifiable, opinion-based content (67 per cent of all posts), often expressing extremist views.

The audit identified AI-generated content as an emerging trend. This included deepfakes of politicians and synthetic avatars (such as cartoon animals) used to disseminate offensive humour, strong political messaging, or hostile commentary. Overall, 5 per cent of political posts were clearly AI-generated.

Findings demonstrated the ongoing deterioration of social media quality, sometimes referred to as 'enshittification', as platforms shift from prioritising users' experience to maximising engagement and monetisation. Bondata's survey indicates that more than one third of respondents encountered problematic content regularly or repeatedly. Bondata's survey also found that half of young respondents reported feelings of disappointment, fear, anger, or sadness when encountering political and social discussions on social media.

Tiivistelmä

Tulevaisuustalo Sitra tilasi Behavioural Insights Teamilta (BIT) ja Bondatalta selvityksen, jossa tutkittiin 18–24-vuotiaiden nuorten sosiaalisessa mediassa kohtaamaa poliittista sisältöä ja arvioitiin sitä, millaisia mahdollisia riskejä siitä aiheutuu kansalaiskeskustelulle. Tällaiset riskit kuuluvat Euroopan unionin digitaalipalvelusäädöksen (DSA) tunnistamiin järjestelmäriskeihin.

BIT toteutti 24 avatarin eli digitaalisen hahmon avulla järjestelmällisen tarkastelun Instagramiin, TikTokiin ja X:ään Suomessa, Ranskassa ja Romaniassa. Avatarit simuloivat eritasoista kiinnostusta poliittiseen sisältöön. Mukana oli myös vaihe, jossa avatarit osoittivat yksinomaan kiinnostusta joko vasemmistolaiseen tai oikeistolaiseen sisältöön. Avatarit kohtasivat kaikkiaan 1 719 poliittista julkaisua, jotka tutkijat luokittelivat. Luokittelussa katsottiin julkaisujen poliittista suuntautumista sekä sitä, havaittiinko niissä ongelmallista sisältöä, kuten virheellisiä väitteitä, salaliittoteorioita tai vihapuhetta.

Bondata täydensi BIT:n tarkastelua verkkokyselyllä, johon vastasi 3 063 nuorta aikuista samoissa maissa. Kyselyssä tarkasteltiin sosiaalisen median käyttöä ja altistumista ongelmalliselle poliittiselle sisällölle.

Oikeistolainen sisältö hallitsi BITin tekemässä tarkastelussa. Sen osuus kaikista poliittiseksi luokitelluista sisällöistä oli 58 prosenttia, kun vasemmistolaista sisältöä oli 26 prosenttia ja keskustaan luokiteltua 16 prosenttia. Oikeistolaisen sisällön hallitsevuus jatkui usein myös tilanteissa, joissa avatarit osoittivat kiinnostusta vasemmistolaiseen politiikkaan. Havainto viittaa oikeistolaisten sisältöjen suhteettomaan vahvistumiseen sosiaalisessa mediassa käyttäjien omista mieltymyksistä riippumatta. Romanian syötteet olivat poikkeus: niitä hallitsi pääasiassa poliittiseen keskustaan luokiteltu sisältö, erityisesti maan hallituksen viestintä. Bondatan kysely tuki pääosin näitä havaintoja. Se osoitti, että 44 prosenttia suomalaisista nuorista aikuisista, jotka määrittelevät itsensä ”vahvasti vasemmistolaisiksi”, koki, että sosiaalisen median sisällöt vastaavat erittäin huonosti heidän näkemyksiään. Vastaava luku ”vahvasti oikeistolaisten” keskuudessa oli vain 5 prosenttia. Ranskassa ja Romaniassa tulokset olivat samankaltaiset.

Tarkastelu paljasti algoritmien arvaamattomuuden ja käyttäjien hallinnan puutteen. Se, minkälaista sisältöä avatarit ilmaisivat suosivansa, ei vaikuttanut niiden näkemään sisältöön johdonmukaisesti eikä luotettavasti. Syötteet saattoivat muuttua äkillisesti ja huomattavasti ilman selvää syytä.

Suurin osa havaitusta ongelmallisesta sisällöstä ei rikkonut alustojen yhteisösääntöjä. Virheelliset väitteet, salaliittoteoriat ja vihapuhe olivat suhteellisen harvinaisia BITin aineistossa. Sen sijaan alustoilla oli runsaasti mielipiteisiin perustuvaa sisältöä, jota ei ole mahdollista tarkistaa (67 prosenttia kaikista sisällöistä) ja jossa usein ilmaistiin äärinäkemyksiä.

Tekoälysisältö on nouseva trendi. Loukkaavan huumorin, jyrkkien poliittisten viestien ja vihamielisten kommenttien levittämiseen käytettiin esimerkiksi poliitikoista tehtyjä deepfake-videoita ja tekaistuja hahmoja (kuten sarjakuvahahmoja). Viisi prosenttia poliittisista sisällöistä oli selvästi tekoälyllä tuotettua.

Tulokset osoittivat somealustojen laadun rapautumisen, kun alustat siirtyvät käyttäjien kokemuksen priorisoinnista sitoutumisen ja kaupallistamisen maksimointiin. Bondatan kyselyssä yli kolmasosa vastaajista kertoi kohtaavansa ongelmallista sisältöä säännöllisesti tai toistuvasti. Puolet vastaajista myös ilmoitti tuntevansa pettymystä, pelkoa, vihaa tai surua kohdatessaan poliittisia ja yhteiskunnallisia keskusteluja sosiaalisessa mediassa.

Sammanfattning

Framtidshuset Sitra gav The Behavioural Insights Team (BIT) och Bondata i uppdrag att undersöka det politiska innehåll som 18–24-åringar möter på sociala medier och att bedöma vilka risker detta kan innebära för den offentliga debatten. Den här typen av risker ingår i de systemiska risker som identifieras i Europeiska unionens förordning om digitala tjänster (Digital Services Act).

BIT genomförde en systematisk undersökning med 24 avatarer på Instagram, TikTok och X i Finland, Frankrike och Rumänien. Avatarerna simulerade olika nivåer av engagemang i politiskt innehåll, däribland exklusivt intresse för antingen vänster- eller högerinriktat innehåll. Totalt kodades 1 719 politiska inlägg manuellt utifrån politisk inriktning och problematiskt innehåll, inklusive desinformation, konspirationsteorier och hatretorik.

Bondata kompletterade undersökningen med en webbenkät bland 3 063 unga vuxna i samma länder, med fokus på användning av sociala medier och exponering för problematiskt politiskt innehåll.

Högerinriktat innehåll dominerade i BIT:s undersökning och utgjorde 58 procent av alla politiskt klassificerade inlägg, jämfört med 26 procent vänsterinriktat och 16 procent mitteninriktat innehåll. Denna dominans kvarstod ofta även när avatarerna signalerade intresse för vänsterpolitik, vilket tyder på en oproportionerlig förstärkning av högerperspektiv oavsett användarpreferenser. Rumänska flöden var ett undantag: de dominerades i stor utsträckning av mitteninriktat innehåll, särskilt myndighetskommunikation. Bondatas enkät bekräftade dessa resultat och visade att 44 procent av användarna i Finland som starkt identifierar sig med vänstern upplevde att innehållet de fick stämde mycket dåligt överens med deras åsikter, jämfört med endast 5 procent av dem som starkt identifierar sig med högern. Resultaten i Frankrike och Rumänien var likartade.

Undersökningen visade på algoritmisk oförutsägbarhet och brist på användarkontroll. Engagemangssignaler hade ingen konsekvent eller tillförlitlig inverkan på vilket innehåll som presenterades. Flöden kunde skifta plötsligt och kraftigt utan en tydlig utlösande faktor.

Majoriteten av det problematiska innehåll som observerades bröt inte mot communityns regler. Renodlad misinformation, konspirationsteorier och hatretorik var relativt ovanliga i BIT:s undersökning. I stället dominerades flödena av icke verifierbart, åsiktsbaserat innehåll (67 procent av alla inlägg), ofta med extremistiska ståndpunkter.

Undersökningen identifierade AI-genererat innehåll som en framväxande trend. Detta inkluderade deepfakes av politiker och syntetiska avatarer (till exempel tecknade djur) som användes för att sprida stötande humor, starka politiska budskap eller fientliga kommentarer. Totalt var 5 procent av de politiska inläggen tydligt AI-genererade.

Resultaten visade på en fortsatt försämring av kvaliteten på sociala medier, vilket ibland kallas för ”enshittification”, i takt med att plattformarna går från att prioritera användarupplevelsen till att maximera engagemanget och intäkterna. Bondatas undersökning visar att mer än en tredjedel av de tillfrågade regelbundet eller upprepade gånger stött på problematiskt innehåll. Bondatas undersökning visade också att hälften av de unga tillfrågade uppgav att de kände besvikelse, rädsla, ilska eller sorg när de stötte på politiska och sociala diskussioner på sociala medier.

1. Introduction: social media and the decline of the digital public sphere

Social media has become the primary source of political and social information for young people in Europe (European Commission 2025b). This shift has coincided with the declining role of traditional media as gatekeepers of information. Social media platforms and their algorithms shape public debate and fuel polarisation, particularly among younger generations. The power of algorithms challenges core democratic principles. While digital platforms have arguably expanded opportunities for free expression, they have not supported democratic values such as respect and constructive dialogue (Dufva et al., 2026).

Online platforms sit at the centre of economic power and are increasingly intertwined with geopolitical competition at a time when the rules-based international order is under strain (Mäkelä et al. 2025). Social cohesion is weakening across societies. Social media has been used in state-led influence operations aimed at deepening divisions and increasing the risk of internal conflict. (Dufva et al. 2026.)

Based on this study, young people spend a staggering amount of time on social media platforms, with daily use averaging over 5 hours in Finland, Romania and France (see *Findings* for further details). High screen time has been linked to increased depressive symptoms and other negative effects on well-being (World Health Organization 2024; Hughes & Borrett 2024).

A growing body of research points to the so-called platform decay or 'enshittification' of social media platforms (Doctorow 2025) – a process in which content quality declines as platforms shift from prioritising users' experiences to maximising engagement and monetisation. Feeds become saturated with sensational, polarising and low-quality material optimised for engagement. For young people, the effects can be particularly harmful.

Artificial intelligence (AI) has long shaped social media through recommendation algorithms. Generative AI has further lowered the threshold for producing content, including text, speech, images, music and video. At the same time, it has made it easier to create and disseminate disinformation, as well as to imitate and replicate authentic content. AI-generated content now blends seamlessly with human-produced material, making it harder to distinguish authentic information from manipulated or fabricated outputs (Dufva et al. 2026). At the same time, certain platforms have scaled back content moderation and replaced third-party fact-checking with user-driven 'community notes', further shifting responsibility to users (Augenstein et al. 2025).

Sitra commissioned the Behavioural Insights Team (BIT) to conduct a systematic platform audit to examine what kind of political content is presented to young users. Bondata complemented this with an online survey exploring young people's experiences of, and emotional responses to, political content on social media. This publication by Sitra presents the main findings from BIT and Bondata, discusses their implications and outlines policy recommendations.

Research questions and methodology

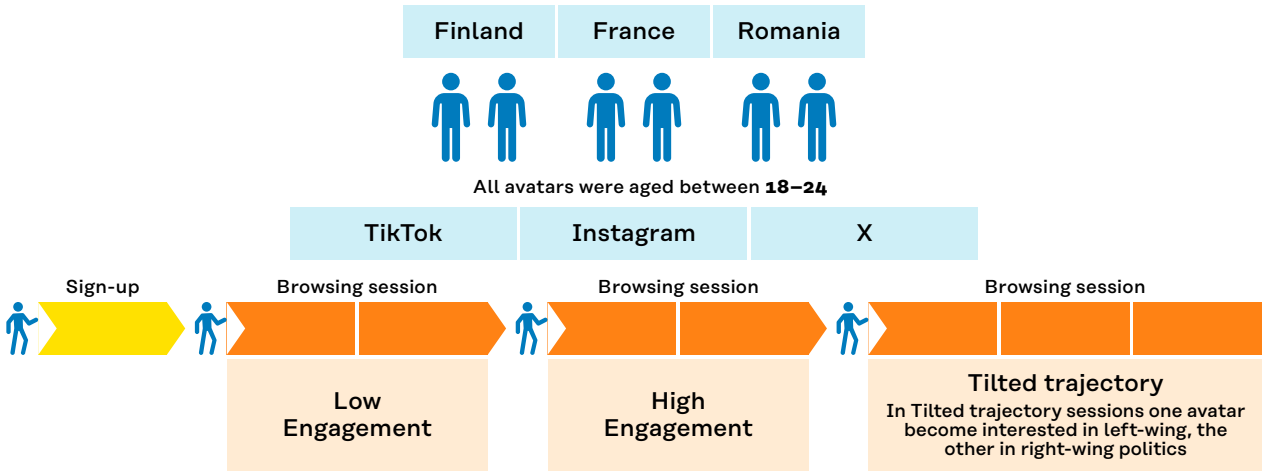
THIS STUDY AIMS TO ANSWER THE FOLLOWING QUESTIONS:

1. What type of political content, paid or unpaid, is presented to young people by the platforms in question and to what extent does this content display ideological imbalance or contain problematic elements (e.g. misinformation, hate speech, or AI-generated material)?
2. How does the political content to which young people are exposed evolve over time as they engage with such material on these platforms and how predictable or responsive are recommender systems to user signals?
3. How do young Europeans themselves perceive the political content they encounter on social media, what emotions does it evoke, and how does it affect their sense of agency and willingness to participate in public debate?

BIT conducted a systematic audit using 24 avatars aged 18–24 on Instagram, TikTok, and X across three EU countries: Finland, France, and Romania. Each avatar completed seven browsing sessions structured into three phases designed to simulate different user behaviours:

1. **Low-Engagement:** Avatars expressed no particular interest in political content.
2. **High-Engagement:** Avatars signalled broad political interest by following major political parties across the political spectrum and spending more time viewing political posts than other content.
3. **Tilted Trajectory:** Avatars signalled exclusive interest in either left-of-centre or right-of-centre content.

Figure 1. Illustration of the scope and methodology of BIT's study.



Researchers manually coded all 1,719 political posts encountered during the audit. Posts were classified by political leaning and assessed for problematic content, including misinformation, malinformation, conspiracy theories, hostile speech, and hate speech.

Bondata, a Finnish research firm, conducted an online survey examining young adults' social media use and their exposure to hate speech, misinformation, and conspiracy theories. The survey was conducted in the same countries as the BIT audit, among young adults aged 18–29, in order to obtain a representative sample. In total, 3,063 responses were collected, evenly distributed across the three countries.

A more detailed overview of the methodology can be found in the Annex of this report, and a full methodology description and descriptions of avatar journeys with selected screenshots can be found in a separate Technical Appendix by BIT.

2. How social media shapes civic engagement based on prior research

2.1 Context: the erosion of democracy and divergent approaches to regulation

A resilient democracy depends on access to trustworthy information and a civic space where misinformation and hate speech do not distort debate. Yet democracy is under increasing pressure. In 2024, 72 per cent of the world's population lived in autocracies – the highest share since 1978 (V-Dem Institute 2025). Although Europe remains largely a bastion of liberal democracy, public support for democracy is not guaranteed. A 2025 study in seven European countries found that only 57 per cent of young people always consider democracy the best form of government, while 21 per cent would accept authoritarian rule under certain circumstances (TUI Stiftung 2025). In Finland, Sitra's Future Barometer shows declining optimism about the future and weakening trust in political decision-making (Rekola et al. 2025).

At the same time, European democracies face external interference and support for anti-democratic movements. According to Eurobarometer 81 per cent of Europeans believe that foreign interference in democratic systems is a serious problem that needs to be addressed. Authoritarian regimes have used social media platforms to seek to erode trust in democratic institutions, and free and fair elections (European Commission 2025a).

In response, governments have strengthened platform regulation, though approaches differ. Democratic jurisdictions such as the EU and the US ground their data governance in democratic values, while authoritarian regimes like China and Russia prioritise state control over data. Within the democratic camp, the EU adopts a rights-based approach emphasising fundamental rights and individual data control, whereas the US takes a more market-oriented view, treating data primarily as an economic asset. (Bradford 2023; Wasastjerna 2020.)

The European Union's landmark initiative in this field is the Digital Services Act (DSA). In addition to harmonising procedures for removing illegal content and increasing transparency, the DSA requires very large online platforms and search engines to assess and mitigate systemic risks stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services. Risk assessments must cover recommender systems, content moderation, terms and conditions and their enforcement, advertising systems and data practices. Presence of illegal content is not a condition for the existence of systemic risks, which are considered as broader societal harms stemming from platform design and functioning. Breaches can lead to fines of up to 6 per cent of global annual turnover. The DSA has also become part of a wider transatlantic political debate as the Trump administration has criticised the DSA as a tool for censorship (Reuters 2026).

2.2. What young people see: political content and bias on social media

Social media platforms have surpassed television as the primary source of political information for young Europeans in 2024. Now 42 per cent of those aged 16–30 cited social media as their primary news source, followed by television (39 per cent) (European Commission 2025b).

The political content encountered on social media platforms has been found in previous studies to be characterised by a disproportionate number of right-wing views. Politico found in 2024 that extremist EU politicians, especially those on the far-right, are more active on TikTok and receive much more engagement than centrists (Goujard et al. 2024). In France, accounts on TikTok expressing right-wing interests quickly led to the creation of echo chambers (Zinigrad 2024). Furthermore, a 2025 journalistic investigation in Germany found that neutral accounts on various platforms saw more right-leaning than left-leaning content, with most of the posts shown supporting the far-right Alternative für Deutschland party (Global Witness, 2025). During the 2024 US elections, Republican content was more often recommended on TikTok (Ibrahim et al. 2025), and new users of X also experienced a right-wing bias in their feeds (Ye et al. 2024).

Content on social media has also been found to often involve misinformation, i.e. statements that are verifiably false: research by BIT following a period of unrest in the UK in the summer of 2024 found that 74 per cent of social media users reported seeing false information on social media in the past week (BIT 2024b). This dynamic is further intensified by the growing presence of AI-generated content: a study analysing over 90,000 posts on X found that AI-generated misleading content is disproportionately more likely to go viral, appears highly realistic, and is just as believable and harmful as traditional forms of misinformation (Drolsbach and Prollochs 2025).

It is not clear, however, whether this overrepresentation of right-wing, and misleading content reflects algorithmic bias, a skewed supply of content by creators, or genuine user preferences. The exact ways in which social media platforms' recommender algorithms work remain unknown. Instagram only provides a high-level explanation of their use of engagement signals, whereby distinct algorithms for different features are deployed, basing ranking on user activity, content metadata, interaction history, and predictive modelling (Mosseri 2021).

A journalistic article reviewing an internal TikTok document suggests that its system relies heavily on watch time and a weighted scoring equation (Smith 2021). The algorithms used by X have been made open source to some extent, revealing details of its four-stage recommendation process – yet many crucial questions about the relative importance of various factors when determining recommendations remain unanswered (QuickFrame 2025; Huertas 2023).

Without a more thorough understanding of what determines the types of content shown to users, it is impossible to know the exact role algorithms play in the spread of problematic or politically skewed content on social media. It is equally important to understand how algorithms balance content shared by users' networks with other content: one study found that the main catalyst for young people's exposure to campaign news is their network of friends and followers, not political accounts (Marquart et al. 2020).

2.3. How social media shapes polarisation and civic discourse

Increased polarisation is a widely studied phenomenon and avenue through which social media might adversely impact civic discourse and overall societal cohesion. The academic literature differentiates between three types of polarisation:

- affective (dislike towards the other group),
- ideological (growing distance between the views of groups),
- social polarisation (preference to limit social contact with the other group), each of which can lower the quality of civic discourse (Arora et al. 2022).

However, the relationship between social media use and polarisation remains complex and contested (Arora et al. 2022; Barberá 2020). While some research indicates that social media contributes to various types of polarisation (Claud 2022; Kubin & von Sikorsky 2021; Levy 2020), other studies present mixed results or even suggest a reduction in polarisation in some cases (Levy 2020; Nyhan et al. 2023; Oden & Porter 2023; Terren & Borge 2021). It is important, therefore, to consider other ways in which social media can pose risks to civil discourse.

Social media use might influence civic discourse by exerting a disproportionate impact on young people's views. Social media influencers can play an important role in shaping young people's political orientation: research in Germany found that influencers increasingly interpret news on their channels (Schmuck et al. 2022) and make politics seem simpler to younger cohorts (Muth & Peter 2023). Young people might also develop a positive psychological relationship with influencers, leading to increased perceived information quality and receptivity to their political messages (Cheng et al. 2023). Furthermore, encountering misinformation can have a disproportionate effect on young people: higher age is associated with a better ability to discern truth from false news (Sultan et al. 2024).

Research has also investigated contextual factors influencing how people engage with news on social media. Studies found that unexpected curation, that is, when the content presented is not what users sought out, prompts users to employ their digital literacy skills and, for example, question the source from which the news comes (Swart 2021 & 2023). The platform of choice for news also matters: one study found that people who saw a news item shared on Facebook were more sceptical than those who saw the same item on a news platform (Karlsen & Aalberg 2023).

It is unclear to what extent and why platform mechanisms amplify certain content and ideologies, and how this differs across users, countries, and platforms. The publicly available knowledge about algorithmic mechanisms is insufficient for independent researchers to accurately model or replicate the user experience. With the implementation of Article 40 of the DSA, vetted researchers will have easier access to such data, providing additional opportunities for future research. Specific algorithmic changes have been tested, such as restricting exposure to like-minded content (Nyhan et al. 2023), using reverse-chronologically-ordered feeds (Guess et al. 2023a), or removing reshared content (Guess et al. 2023b). However, their long-term effectiveness in reducing polarisation or shifting beliefs is limited, suggesting that deeply held attitudes and the broader information ecosystem may exert greater influence than platform mechanics.

Why users spend so much time online

Social media use is shaped by psychological tendencies, platform design, and the way

information is presented. While some of the time spent on social media reflects genuine enjoyment, many users report wishing they – and others – used these platforms less (Bursztyn et al. 2023). Research suggests that this pattern is not simply a matter of preference, but the result of behavioural mechanisms and platform design. Some key psychological drivers of spending time on social media are presented in the figure below. (BIT 2025; Flayelle et al. 2023; Bursztyn et al. 2023; Berger & Milkman 2012; Laibson 1997; Thaler & Shefrin 1981)

Figure 2. Psychological drivers of spending time on social media.

Collective traps	Choice architecture	Emotional content	Inconsistent preferences
Users continue using social media although they would **prefer a world in which it did not exist.** This dynamic is known as a "collective trap", where opting out individually carries a social cost.	**Design elements can encourage prolonged use,** such as infinite scroll, likes and social rewards, personalised feeds, and temporary content that creates fear of missing out.	**Emotionally charged content,** like hate or conspiracies, can **dominate feeds due to high engagement,** regardless of how much users agree with it.	People often **plan to limit their social media use,** but **present bias** leads people to act against their longer-term goals. This is known as **time-inconsistency**.

Defaults are among the most powerful tools in online choice architecture (OCA). When options are pre-selected, most users stick with them, not because they reflect considered preferences, but because defaults reduce cognitive effort and signal recommended behaviour (Thaler & Sunstein 2021; Thaler & Shefrin 1981; Laibson 1997). On social media, defaults strongly shape time spent, content exposure and notification intensity. Small barriers, or 'sludge', can further discourage users from adjusting settings, undermining meaningful choice and trust (Thaler 2018).

Evidence shows that OCA can be used to promote better outcomes. Simple design changes – such as clearer prompts and simplified language – significantly improve engagement with terms, conditions and content controls (BIT 2019; BIT 2024a; BIT 2024d). Academic research also demonstrates that accuracy prompts reduce the sharing of misinformation (Pennycook & Rand 2022).

The key structural problem in the social media is 'shrouding': critical information about recommender systems, content moderation and user controls is often hidden, difficult to interpret or hard to compare (BIT 2024c). This obscurity limits users' ability to understand why certain content appears in their feeds and weakens incentives for platforms to compete on safety or democratic quality. Greater transparency, particularly regarding amplification logic and control tools, is essential to restore meaningful agency. However, transparency alone is insufficient if information remains complex. Research shows that standardised comparability tools, such as clear labels or benchmarks, help users navigate difficult choices and encourage providers to compete on what matters most (Loewenstein et al. 2014).

Research findings suggest that platform design can either exploit behavioural biases for engagement or be deliberately structured to support informed, responsible participation.

Why exposure to problematic content online shapes attitudes

Behavioural science can also help us understand how exposure to problematic content, such as misinformation, hate speech, conspiracy theories, or extremist materials can shape attitudes, beliefs, and civic discourse.

Some key known mechanisms are illustrated in the figure below (Mackinac Center n.d.; Kubin & von Sikorski 2021; Zajonc 1968; Muth & Peter 2023; Cheng et al. 2024; Madriaza et al. 2025).

Figure 3. How problematic content shapes attitudes and discourse.

Overton window	Exposure effect	Influencers	Social norms	Mental health
Exposure to extremist views can shift what is viewed as acceptable, making **fringe opinions seem mainstream**.	People tend to **prefer ideas they encounter frequently**, regardless of argument strength. Seeing extremist content repeatedly can make it seem more persuasive.	People often form **parasocial relationships** with influencers, who deliver **simplified political messages** in relatable formats. These processes allow influencers to relay their political views effectively.	Observing how others debate political issues online can **shape users' own communication styles** both online and offline.	Repeated exposure to hostile or extremist content has also been found to **negatively affect wellbeing and normalise hate**.

Next, the key findings of the study are presented.

3. Findings: right-wing bias, memes and lack of control

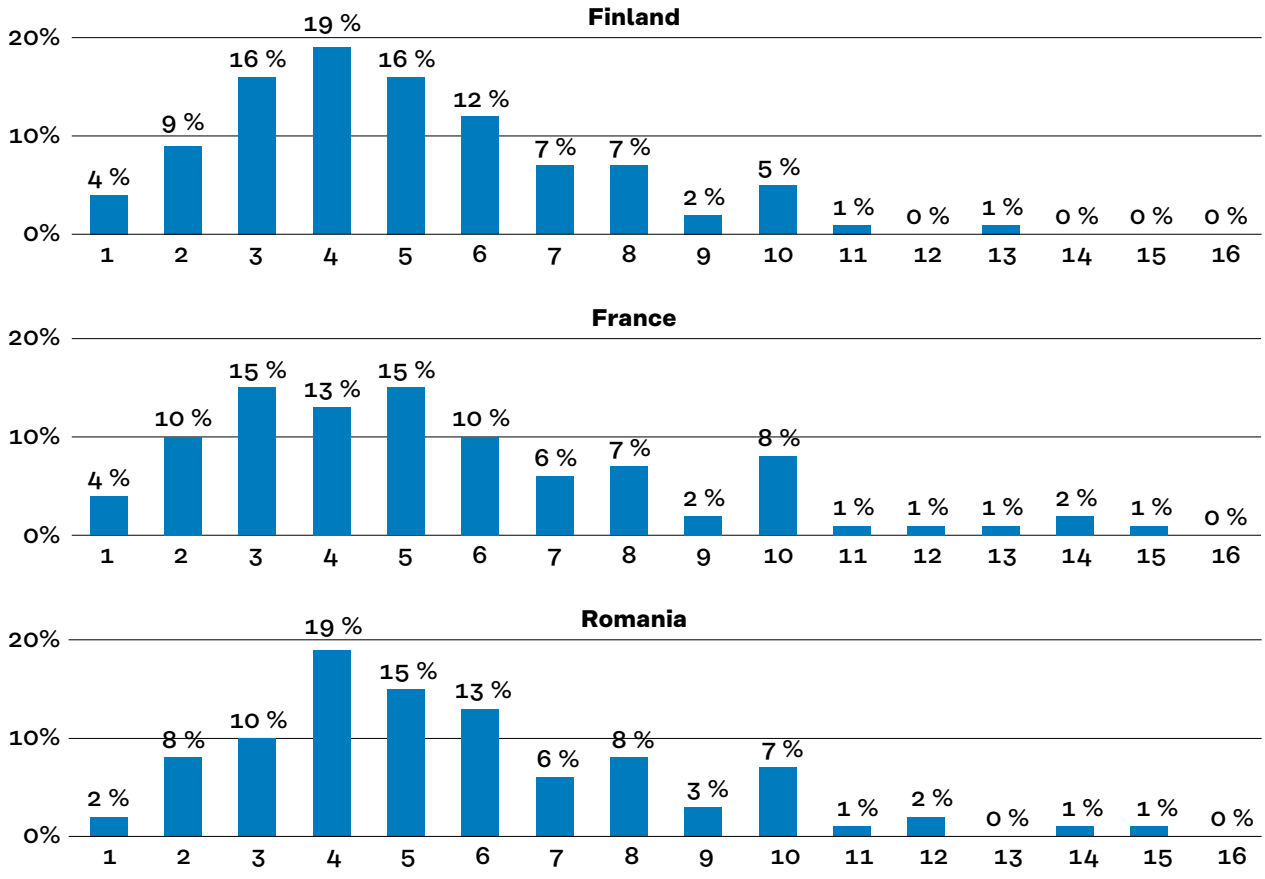
3.1. Overview of the datasets

The BIT platform audit was carried out in two phases. The first audit covered all three platforms and countries after which a second audit was conducted to verify the results of the first audit, covering the three platforms, but only focusing on Finland. Across the three platforms and countries, the first audit identified 1,337 political posts. Of these, 718 appeared on X, 379 on TikTok, and 240 on Instagram. By geography, 512 were seen by Finnish avatars, 415 by French avatars, and 410 by Romanian avatars.

The second audit identified 382 political posts in total in Finland. Of these, 218 appeared on X, 108 on TikTok, and 56 on Instagram.

Each dataset reflects seven browsing sessions per avatar, capturing both algorithmic recommendations and posts from followed accounts. It is worth noting that the researchers did not identify any paid political advertising during the journeys.

Figure 4. Time spent on social media by 18–24-year-olds in Finland, France and Romania. Source: Bondata 2025.



More information, including a glossary of terms used, can be found in the Annex.

Bondata's survey was conducted in the same countries as the BIT audit, but with a wider age range (18–29) to obtain a more representative sample. Survey data consists of 3,063 responses. Response distribution by country is as follows: Finland (1,030), France (1,022) and Romania (1,011). The study's margin of error at the 95 per cent confidence level is ±3.1 percentage points.

3.2. Daily use of social media by young adults

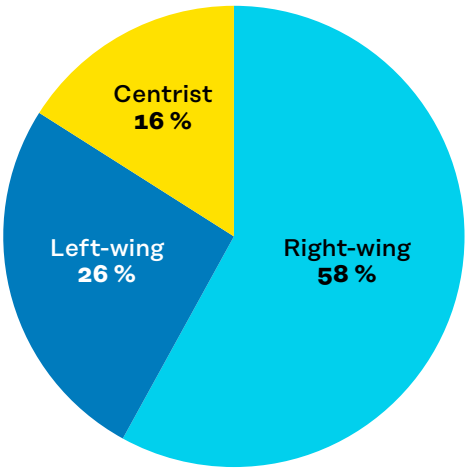
According to Bondata's survey, daily social media use averages for 18–24-year-olds are 5.3 hours in Finland, 6.2 hours in France and 6.1 hours in Romania (Bondata 2025). On average, almost one in four young respondents reported spending at least 8 hours per day on social media.

3.3. Prevalence of right-wing, left-wing, and centrist content

Expressing political views across the full political spectrum is a fundamental part of democratic debate and freedom of expression. BIT dataset represents a wide range of political views and topics. Despite this, individual browsing sessions were often dominated by a single partisan or ideological viewpoint, even when the avatar had not yet indicated a left- or right-wing political interest.

Overall, right-wing content was by far the most prevalent across the dataset.[1] Of 1,151 posts across the two audits that could be politically classified as right-wing, left-wing, or centrist, 58 per cent were right-wing, 26 per cent left-wing, and 16 per cent centrist.

Figure 5. Overall prevalence of right-wing, left-wing, and centrist content in three countries, based on two audits conducted by BIT.



Right-wing content was, on average, more prevalent even for the sessions where the avatars had already expressed an interest in left-wing politics (12 sub-journeys across the two audits, involving 36 browsing

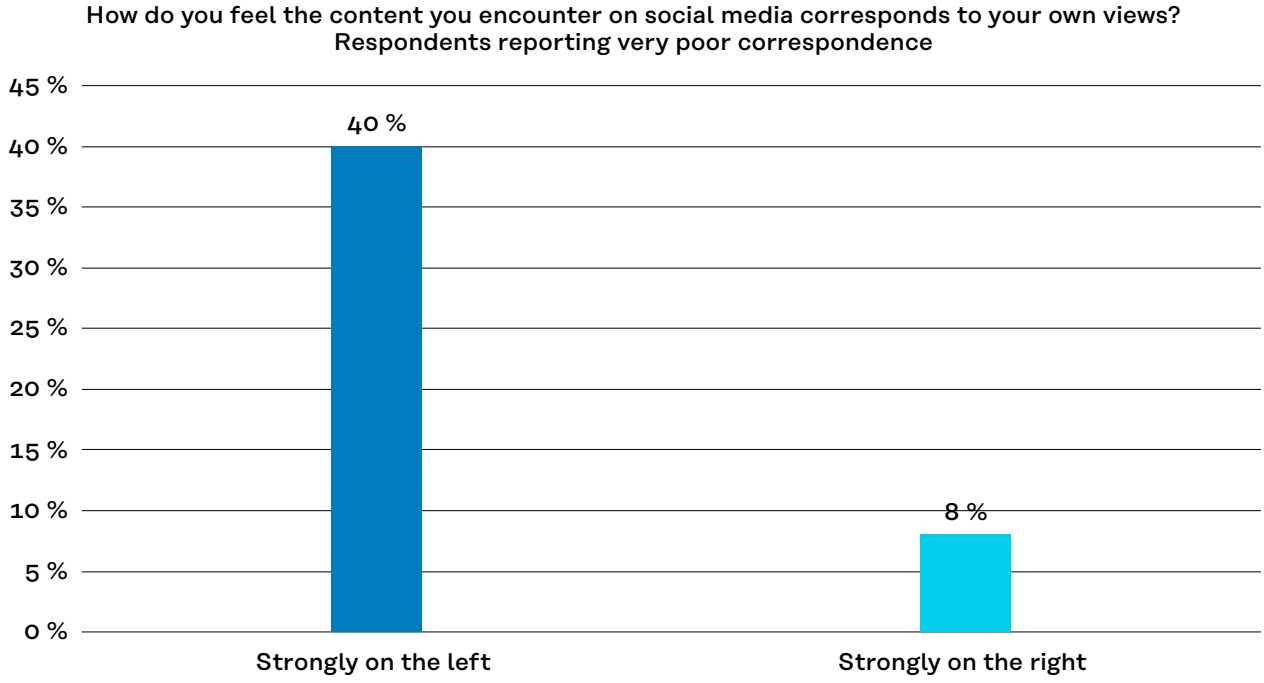
1 Example posts demonstrating how political posts were categorised:

- Right-wing: A post showing Jordan Bardella (leader of the right-wing National Rally party in France) making speeches. An AI-generated voiceover praises him and the caption reads 'Here is Jordan Bardella, symbol of unbounded patriotism'. (France, TikTok)
- Left-wing: A video by the Finnish Trade Union Confederation about why they feel vocational education is broken. (Finland, TikTok)
- Centrist: A video of a centrist Romanian minister accusing AUR (right-wing party) of always complaining but never proposing any real actions. The caption says that this is a real minister, not the "the dogs" from PSD (social-democratic party, with traditionalist views on social issues). (Romania, X)
- Could not be categorised: A post saying its creator is happy that Ion Iliescu (former president of Romania) is ill, and wishes him agony. (Romania, X). It was unclear why the creator is hostile towards Iliescu.
- Neutral: A video explaining new taxes in France without passing a judgement or expressing an opinion on the taxes introduced. (France, Instagram).

sessions and 388 political posts). In these sessions, the avatars still saw 182 right-wing posts compared with 59 left-wing and 43 centrist posts. Overall, out of the 12 sub-journeys, seven had more right-wing than left-wing posts. This means that, on average, sending signals about an interest in left-wing politics did not lead to a higher proportion of left-wing content in the feeds; instead, right-wing views continued to dominate.[2]

At the level of the whole sample Bondata's survey indicates no perceived bias on political content: 29 per cent of respondents considered their feeds as leaning left and 25 per cent as leaning right.[3] However, when examining the political orientation of the respondents, the Bondata survey substantiated BIT's finding of the prominence of right-wing content for left leaning users. In Finland, 44 per cent of users who identified themselves 'strongly on the left' felt that the content they received corresponded very poorly to their own views. Only 5 per cent of users 'strongly on the right' reported the same. In France and Romania results followed a similar pattern: 38 per cent in France and 37 per cent in Romania of those 'strongly on the left' and

Figure 6. Social media users who felt that the content they received corresponded very poorly to their own views, total in three countries. Source: Bondata 2025.



2 The distribution of the high number of right-wing posts in the BIT audit was uneven across countries, platforms, and browsing journeys. Therefore, some journeys were dominated by left-wing content instead, or offered a balance between right-wing, centrist, and left-wing views. The reason behind this variation remains mostly unclear, as many changes in the feeds did not reflect the avatars' signaled preferences.

3 This difference from the audit findings likely reflects that surveyed users have long-standing, personalised feeds, whereas the audit used new accounts, as well as potential gaps between users' subjective perceptions and researchers' classifications of political content. BIT's platform audit using avatar accounts measured observed content under controlled conditions while Bondata's survey measured self-reported perceptions among real users. This difference in methods means that comparability between the two data-sets is limited.

11 per cent in France and 10 per cent in Romania of those 'strongly on the right' felt social media content corresponded very poorly to their own perspectives. Across the three platforms an average of 40 per cent of those 'strongly on the left' and 8 per cent of those 'strongly on the right' felt that content they received corresponded very poorly to their own views.

This asymmetry persisted in all three countries. While seeing opposing views on social media is valuable and part of a pluralistic debate, people strongly on the left and strongly on the right differ sharply in how well they feel the content matches their own views.

3.4. Political classification of content across countries

In Finland, we observed an overall right-wing bias, but there was a strong variation across platforms and the two audits, which could not be explained by the signals the avatars sent. Notably, the Instagram journeys in the first audit were dominated by right-wing posts, but in the second audit, this trend flipped and left-wing views dominated. In both audits, TikTok offered a balance across the political spectrum while X had a clear right-wing bias.

On average, France displayed a more balanced distribution of posts across the political spectrum. There was a higher share of left-leaning posts on X and a right-wing bias on Instagram and TikTok.

Romania stood out for its centrist bias on TikTok, largely driven by official communications from the governing party and its politicians. Centrist content was also prominent on X, almost equalling the number of right-wing posts. Instagram displayed very little political content in Romania.

There were political posts that were political in nature but did not express opinions or could not be clearly categorised as left-wing, centrist, or right-wing. These posts were omitted from this analysis but included in all subsequent analyses.

Table 1: Breakdown of posts by political classification across the 'Low-Engagement', 'High-Engagement', and 'Tilted Trajectory' phases

First audit			
Platforms	**Right-wing**	**Left-wing**	**Centrist**
Instagram	135 (81%)	28 (17%)	4 (2%)
TikTok	132 (51%)	69 (27%)	58 (22%)
X	243 (55%)	130 (29%)	69 (16%)
Countries	**Right-wing**	**Left-wing**	**Centrist**
Finland	308 (75%)	86 (21%)	14 (3%)
France	130 (51%)	102 (40%)	24 (9%)
Romania	72 (35%)	39 (19%)	93 (46%)
Total	**510**	**227**	**131**
Total as % of all political posts	38%	17%	10%
Total as % of all political posts that could be categorised as right/left/centre	59%	26%	15%
Second audit (Finland only)			
Platforms	**Right-wing**	**Left-wing**	**Centrist**
Instagram	0 (0%)	22 (58%)	16 (42%)
TikTok	39 (48%)	38 (46%)	5 (6%)
X	119 (73%)	15 (9%)	29 (18%)
Total	**158**	**75**	**50**
Total as % of all political posts	41%	20%	13%
Total as % of all political posts that could be categorised as right/left/centre	56%	27%	18%
Total audits 1&2	**668**	**302**	**181**
Total audits 1&2 as % of all political posts	39%	18%	11%
Total audits 1&2 as % of all political posts that could be categorised as right/left/centre	58%	26%	16%

3.5. Understanding problematic political content online

In BIT's platform audit, only a minority of political posts met the threshold for various types of problematic content (such as misinformation or hate speech) across countries, platforms and the two audits, as outlined in Tables 3a and 3b. The definitions of categories of problematic content are provided in the Annex of this report.

Instead, opinion-based content dominated, which by its nature is unverifiable. Of the 1,719 political posts identified across the dataset, only 455 (26 per cent) could be fact-checked and categorised as true or false, while 1,151 (67 per cent) were opinion or entertainment pieces expressing personal views on political events. Furthermore, 113 (7 per cent) contained unverifiable claims, such as videos alleging criminal acts by specific groups without evidence.

Of the 455 fact-checkable posts, only 54 (12 per cent) were identified as misinformation and 42 (9 per cent) as malinformation, with 359 (79 per cent) confirmed as verifiably true. Across all the political posts identified, misinformation represented roughly 3 per cent of the content.

Opinion-based content was often extremist but typically fell just short of violating platform community guidelines. For example, the avatars encountered hostile posts that, while not containing misinformation, featured derogatory content. Examples include AI-generated videos of a gorilla making misogynistic and xenophobic jokes using swear words, or memes using fictional Nazi characters from popular films to implicitly convey extremist views towards minorities. In this dataset, extremist content was predominantly far-right in ideological orientation and was consequently coded as right-wing. This should not be interpreted as implying that mainstream right-wing political views are extremist.

Such a post might not go against the platforms' community guidelines, but repeated exposure to similar content can have a corrosive impact on civic discourse. It is possible that content creators posted more content violating community guidelines, but the avatars rarely encountered such content due to content moderation.

Table 2a: Breakdown of posts based on accuracy of content by country.

	Mis-information	Mal-information	Unverifiable statements	Opinion and entertainment	True information
Audit 1					
Finland	15 (3%)	8 (2%)	13 (3%)	419 (82%)	57 (11%)
France	8 (2%)	10 (2%)	18 (4%)	252 (61%)	127 (31%)
Romania	24 (6%)	21 (5%)	71 (17%)	184 (45%)	110 (27%)
Total Audit 1	**47**	**39**	**102**	**855**	**294**
Total as % of all audit 1	4%	3%	8%	64%	22%
Audit 2					
(Total) Finland	7	3	11	296	65
Total as % of all audit 2	**2%**	**1%**	**3%**	**77%**	**17%**
Total Audit 1 and 2	**54**	**42**	**113**	**1,151**	**359**
Audit 1 and 2 total as a % of all political posts	3%	2%	7%	67%	21%

Table 2b: Breakdown of posts based on accuracy of content by platform.

	Mis-information	Mal-information	Unverifiable statements	Opinion and entertainment	True information
Audit 1					
Instagram	10 (4%)	4 (2%)	7 (3%)	205 (85%)	14 (6%)
TikTok	3 (1%)	7 (2%)	31 (8%)	253 (67%)	85 (22%)
X	34 (5%)	28 (4%)	64 (9%)	397 (55%)	195 (27%)
Total Audit 1	**47**	**39**	**102**	**855**	**294**
Total as % of all political posts Audit 1	3%	3%	8%	64%	22%
Audit 2					
Instagram	0 (0%)	0 (0%)	0 (0%)	37 (66%)	19 (34%)
TikTok	1 (1%)	1 (1%)	1 (1%)	85 (79%)	20 (19%)
X	6 (3%)	2 (1%)	10 (5%)	174 (80%)	26 (12%)
Total Audit 2	**7**	**3**	**11**	**296**	**65**
Total as % of all political posts Audit 2	2%	1%	3%	77%	17%
Total Audits 1&2	**54**	**42**	**113**	**1,151**	**359**
Total as % of all political posts Audits 1&2	3%	2%	7%	67%	21%

Table 3a and 3b: Breakdown of conspiracy theories

	Number of posts containing a conspiracy theory
Audit 1	
Finland	16 (3%)
France	26 (6%)
Romania	23 (6%)
Total	**65**
Total as % of all political posts	5%
Audit 2	
Finland	2
-	-
-	-
Total	**2**
Total as % of all political posts	1%
Total audit 1&2	**67**
Total as % of all political posts Audits 1&2	4%

	Number of posts containing a conspiracy theory
Audit 1	
Instagram	22 (9%)
TikTok	10 (3%)
X	33 (5%)
Total	**65**
Total as % of all political posts	5%
Audit 2	
Instagram	0 (0%)
TikTok	0 (0%)
X	2 (1%)
Total	**2**
Total as % of all political posts	1%
Total audit 1&2	**67**
Total as % of all political posts Audits 1&2	4%

Table 4a and 4b: Breakdown of posts containing hate speech or hostile speech

	Number of posts containing hate speech	Number of posts containing hostile speech
Audit 1		
Finland	9 (2%)	42 (8%)
France	3 (1%)	9 (2%)
Romania	1 (<1%)	9 (2%)
Total	**13**	**60**
Total as % of all political posts	1%	4%
Audit 2		
Finland	0 (0%)	10 (3%)
-	-	-
-	-	-
Total	**0**	**10**
Total as % of all political posts	0%	3%
Total audit 1&2	**13**	**70**
Total as % of all political posts	1%	4%

	Number of posts containing hate speech	Number of posts containing hostile speech
Audit 1		
Instagram	2 (1%)	22 (9%)
TikTok	4 (1%)	15 (4%)
X	7 (1%)	23 (3%)
Total	**13**	**60**
Total as % of all political posts	1%	4%
Audit 2		
Instagram	0 (0%)	0 (0%)
TikTok	0 (0%)	1 (1%)
X	0 (0%)	9 (4%)
Total	**0**	**10**
Total as % of all political posts	0%	3%
Total audit 1&2	**13**	**70**
Total as % of all political posts	1%	4%

While the BIT platform audit identified relatively few posts that clearly met defined categories of problematic content, Bondata's survey indicates that young people encounter such content regularly. This difference may partly reflect the fact that survey respondents may apply different definitions or thresholds for hate speech, hostile speech, misinformation, and conspiracy theories than those used in the BIT platform audit.

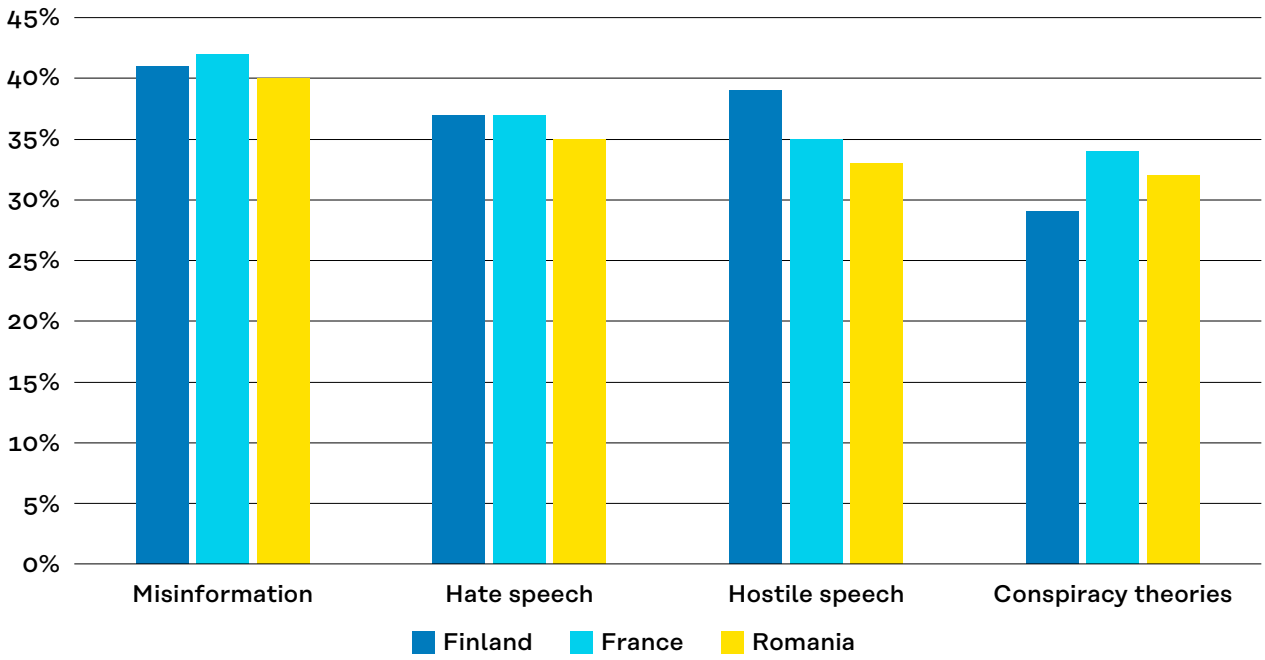
Across the three countries, 37 per cent of respondents reported seeing **hate speech** – content promoting hatred or violence against specific groups – repeatedly or often. The shares were identical in Finland (37 per cent) and France (37 per cent), and only slightly lower in Romania (35 per cent).

Hostile speech is content that degrades, offends or discriminates against specific groups and was also widely reported. It was encountered regularly or repeatedly by 39 per cent of respondents in Finland, 35 per cent in France and 33 per cent in Romania.

Misinformation is likewise perceived to be common. In Finland, 41 per cent of respondents said they regularly or repeatedly encountered content that later proved to be incorrect or misleading. The corresponding figures were 42 per cent in France and 40 per cent in Romania.

Conspiracy theories are defined as content that suggests that significant events or situations are secretly controlled by powerful actors and were encountered regularly or repeatedly by 29 per cent of respondents in Finland, 34 per cent in France and 32 per cent in Romania.

Figure 7. Problematic content encountered regularly or repeatedly. Source: Bondata 2025.



3.6. The presence of AI-generated political content

Overall, 90 political posts (5 per cent) were flagged by the researchers as AI-generated, and there were 39 additional posts (2 per cent), where the researchers suspected that AI might have been used in content generation. There were three main types of AI-generated political content:

1. **Deep fakes.** These were videos (or pictures accompanied by audio) that mimicked well-known political figures. For example, there was a video showing Donald Trump physically abusing Greta Thunberg.
2. **AI-generated political meme videos.** These often-depicted well-known political figures but were clearly intended for entertainment. For example, there was a video showing Emmanuel Macron and other French politicians posing as rappers.
3. **AI-generated humans or animals sharing political opinions.** These were videos in which AI-generated humans (ranging from newborn babies to elderly protesters) or AI-generated animals shared strong political opinions. For example, the avatars saw multiple videos of gorillas making offensive, sometimes hateful jokes about minorities.

In both audits, AI-generated content was overwhelmingly right-wing and often contained hostile speech or offensive mockery. Of the 90 AI-generated posts observed 56 (61 per cent) were right-wing and just 6 left-wing (7 per cent). 16 (18 per cent) contained hostile speech and 2 (2 per cent) hate speech. The remaining seven posts were not classified as hostile speech but were still often offensive due to strong language, malicious intent, or mockery of ethnic groups.

3.7. Algorithmic unpredictability

Engagement signals did not reliably influence content recommendations. Yet, the composition of content often changed suddenly and substantially, without any clear trigger. For instance, one of the Finnish avatars encountered no political content on Instagram in six of the seven browsing sessions. In the final session, however, the feed became dominated by a specific form of political content, notably extreme right-wing memes. These included memes using fictional Nazi characters, most frequently Hans Landa from the 2009 movie 'Inglourious Basterds', to implicitly suggest support for Nazism or other racist ideologies.

This contrasts with the second audit, in which the Finnish Instagram avatar was neither exposed to right-leaning, extremist, or hostile content. Instead, the left-leaning avatar was shown predominantly left-leaning material, while the right-leaning avatar was primarily exposed to centrist content.

Taken together, these findings underscore the unpredictability of algorithmic content delivery.

3.8. What political content looks like across platforms and countries

In BIT's platform audit, the distinctive national news stories in each country led to a variation in the themes of the posts across the audited platforms. In the second audit, despite the importance of local politics, the avatars also saw topical international content, primarily discussing US politics, the Russia-Ukraine and the Israel-Palestine conflicts, and discussions about the Bondi Beach gunman attack. Some of these posts were in English, shared by internationally recognised figures or outlets.

Some journeys were dominated by posts that were primarily jokes, albeit with political content. These were classified as parodies and memes to distinguish them from more substantial political commentary, some of which also used humorous elements, but to a lesser extent.

During the second audit, a scandal involving a Miss Finland titleholder featured prominently across all three platforms in Finland, appearing in posts from across the political spectrum. The controversy arose after an image of her circulated on social media and was interpreted as mocking Chinese people.

How Political Content differs Between Finland, France, and Romania

This section outlines the main similarities and differences across the three audited countries, acknowledging the limits of cross-country comparison given differing political contexts and norms, while highlighting which findings are country-specific and which are shared.

While right-wing content dominated overall, there were a few exceptions. Romanian avatars saw more centrist content (93 posts) than right-wing (72) or left-wing (39) content, driven by videos shared by the centrist party in government and its politicians. The journeys on X in France also displayed more left-wing content (59 posts) than right-wing (31) or centrist (19) content overall. In the second audit, Instagram did not show any right-wing content to the Finnish avatars, in stark contrast to the experience on the same platform during the first audit.

Furthermore, different kinds of problematic content were prominent in the three countries: Finnish avatars were exposed to the most hate and hostile speech, driven by the content seen in the first audit. French avatars saw the most conspiracy theories, and Romanian avatars saw the most misinformation and malinformation posts.

The themes of the posts differed across the three countries. Posts with country-specific themes were the most popular across all three countries. This catch-all theme included reactions to daily news events, discussions of country-specific political issues, and personal opinions on various national politicians. In Finland, other popular themes included immigration and refugees, parodies and memes and the Israel-Palestine conflict. In France, posts centred on the Israel-Palestine conflict, and parodies and memes were also popular, alongside posts focusing on social commentary, such as reflections on cultural issues. In Romania, country specific issues were by far the most common, followed by misogynistic posts, historical discussions, and the Russia-Ukraine conflict.

Finally, the experience of browsing Instagram was fundamentally different across the three countries. In Romania, the avatars saw little political content on Instagram and only a small subset of this content was problematic. In France, the avatars saw the largest amount of conspiracy theories on Instagram. In Finland, however, there was a large amount of hostile and hateful political content on Instagram in the first audit, mostly using meme formats. Most posts advertised Nazism and racism as acceptable and popular ideologies or expressed hatred towards Jewish and black people, sexual minorities, and women. Even those posts not categorised as hate speech or hostile content were distasteful, for example, posts making fun of the Holocaust and its victims. One of the browsing sessions in the 'Tilted trajectory' phase was fully dominated by such content.

How Instagram, TikTok, and X deliver political content

This section discusses key differences and commonalities between the three audited platforms. This comparison is limited by the inherent differences in platforms and the design of their feeds. For example, Instagram's Reels feed is designed to show primarily new content to the user, limiting exposure to the accounts followed, while TikTok's and X's main feeds display content by both followed and recommended accounts. There are also differences in the format of the content displayed. For example, X displays a mix of video-, text-, and picture-based content, while Instagram and TikTok favour video content.

X had by far the most political content across the dataset (936 posts compared to 487 on TikTok and 296 on Instagram). There were some browsing sessions on TikTok and Instagram with no or barely any political posts seen. In comparison, all sessions on X produced a large amount of political content. TikTok, and to a lesser extent Instagram, featured a notable theme of AI-generated content. This type of content appeared harmless at first – for example, AI-generated images of gorillas pouring tea – but it was paired with harmful language, including racist jokes.

This also contributed to the relative predictability of X's feed compared to the other two platforms: across browsing sessions and journeys, the content shown on X did not change dramatically. By contrast, there were sudden spikes in the volume of political content on TikTok and Instagram, or unexplained changes in the characteristics of the political content shown. One notable example was a Finnish journey on Instagram in the first audit, where the first six browsing sessions had no political content, but then, without any obvious trigger, the seventh journey was full of extremist memes.

Finally, in the first audit there was some variation in the themes shown across the platforms for the three countries. On Instagram, there were 52 posts classified as parodies or memes, making them the most popular content type, whereas on TikTok and X country-specific issues dominated with 159 and 252 posts respectively. Popular themes on Instagram included social commentary, country-specific issues, Nazism/Holocaust, and various conspiracy theories (that is, conspiracy theories not captured by other theme categories). On TikTok, other common themes were parodies and memes, immigration and refugees, and posts which were focused on racism or ethnic minorities. On X, popular themes were immigration and refugees, the Israel-Palestine conflict, and US politics.

3.9. What we learned about political content on social media

In BIT's platform audit, across all countries and platforms, the audited feeds were characterised by ideological imbalance, lack of transparency, and instability. While clear misinformation was relatively rare, most of the political content blurred the line between factual reporting and personal commentary: most of the posts presented opinions, often extremist, without referring to verifiable facts. Notably, the avatars were also exposed to AI-generated content used to mimic politicians or express strong political opinions. The presence of AI-generated content was particularly worrying in the second audit, when most of such content was labelled as hostile speech, misinformation, or simply had a malicious message. With the spread and improvement of this technology, this type of content is likely to become more pervasive and harder to identify.

Right-wing content was overrepresented, even in browsing sessions curated through neutral or left-leaning behaviour. This imbalance was the most consistent on X. On other platforms, there were some journeys where content was

balanced across the political spectrum or had a left-wing bias. This was, however, occasional and less frequent than feeds with a right-wing bias. These findings suggest that users are often presented with ideologically siloed feeds, which might not be aligned with their interests.

At the same time, **algorithms behaved unpredictably**. Engagement cues – such as following or unfollowing political parties, liking their content or spending more time on certain types of posts – had no consistent impact on what avatars saw. In many cases, feeds shifted dramatically without any apparent reason, exposing users to sudden surges of extreme or partisan material.

Some of the extremist posts documented in the audit clearly violated platform community guidelines. However, more typically, their creators used **humour, vagueness, or other rhetorical devices** to avoid detection and removal. These types of posts – often styled as memes or parody – were prominent across all platforms and countries, subtly normalising hostility towards minorities and political opponents.

Taken together, these findings indicate that political exposure on social media is neither transparent nor reliably shaped by user behaviour. Instead, it reflects a complex interplay of **algorithmic curation, platform design, and local political context** – one that leaves users vulnerable to potentially one-sided, often emotionally charged, and sometimes extremist narratives, with limited ability to control what they see, or understand why they are seeing it.

3.10. How did social media users perceive political content, and how did it affect their agency?

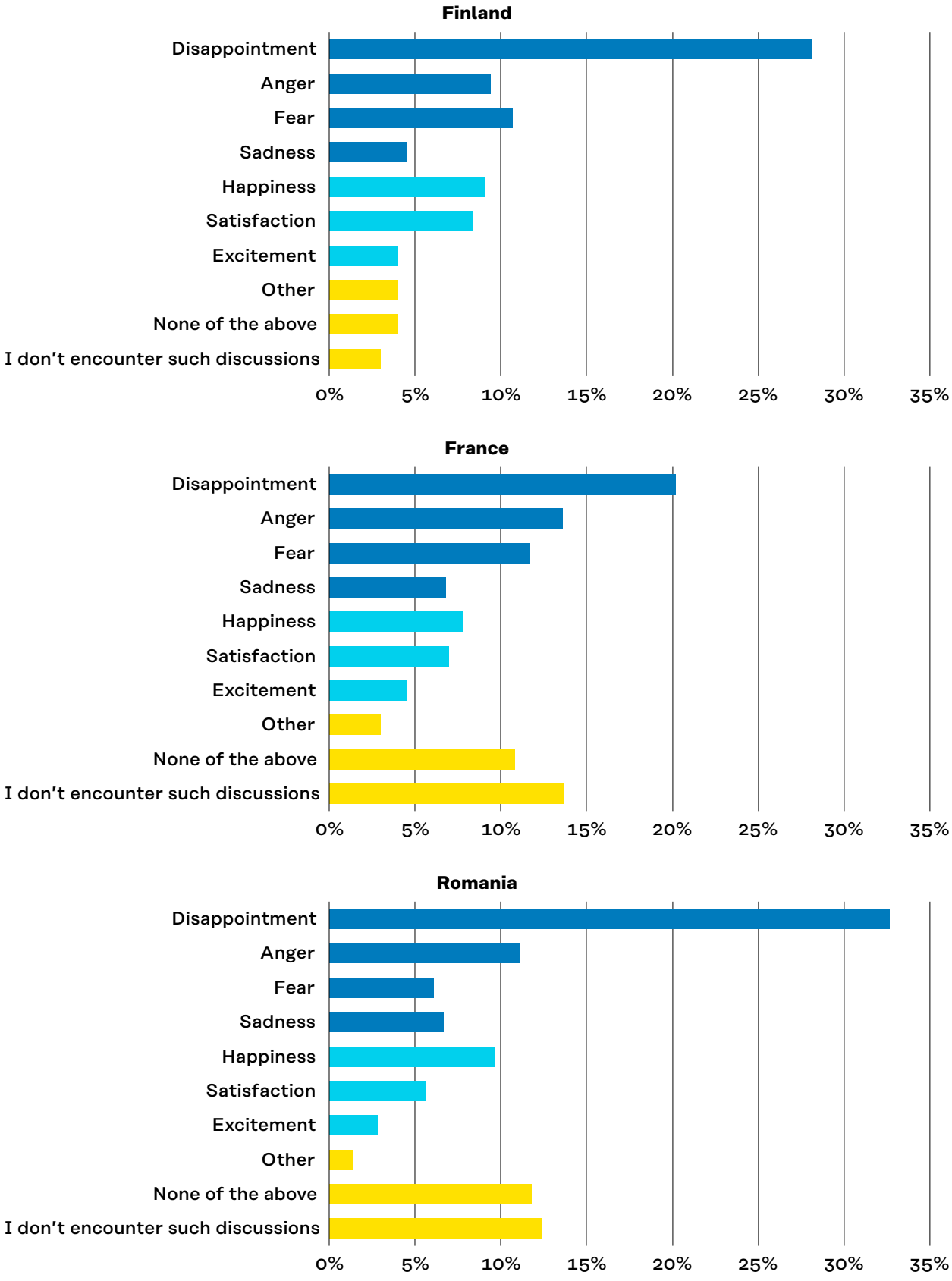
According to the Bondata survey, when users encountered discussions on politics and social issues on social media, the most common resulting emotion across all three countries was disappointment: Finland 28 per cent, France 20 per cent, and Romania 33 per cent. Overall, as many as half of the young people surveyed reported experiencing some form of negative emotion in such situations: disappointment (27 per cent), anger (11 per cent), fear (10 per cent), and sadness (6 per cent). By contrast, a considerably smaller share reported positive emotions – happiness (9 per cent), satisfaction (7 per cent), or excitement (4 per cent) – amounting to only around one in five.

The survey data revealed notable gender differences in emotional responses to political content on social media, particularly in Finland. In this country, 58 per cent of women reported experiencing negative emotions when encountering political discussions, compared with 37 per cent of men – a gap of 21 percentage points, while women were also more likely to feel concerned about humour conveying problematic messages and to report reduced optimism about societal development. By contrast, gender gaps were considerably smaller in France (54 per cent of women vs. 51 per cent of men) and Romania (60 per cent vs. 53 per cent).

For young people, social media functions not only as a source of political information but also as a space for civic participation. Although a majority reported negative emotions in response to political discussions, 41 per cent overall said such content at least somewhat increases their willingness to engage in public debate – a pattern consistent across Finland and France (40 per cent) and slightly higher in Romania (43 per cent).

At the same time, 17 per cent felt that political content reduced their willingness to participate. Men (44 per cent) were more likely than women (39 per cent) to report that it strengthens their engagement.

Figure 8. Feelings raised by discussion on political and social issues on social media in Finland, France and Romania. Source: Bondata.



4. Discussion: how social media undermines civic discourse – and what to do about it

Drawing on the academic research, BIT's platform audit and the Bondata-survey, seven key conclusions are outlined below.

1. Problematic content on social media evades fact-checking and degrades civic discourse

Findings show that out of 1,719 political posts encountered by avatars on social media, very few could be categorised as factually false or misleading, because most of the posts did not make any fact-checkable statements. Instead, two thirds of political posts consisted of opinion or entertainment content, much of it unverifiable and frequently extremist in tone. Memes, irony and humour often function as vehicles for ideological signalling while evading moderation thresholds.

Survey data from Bondata indicates that more than one third of respondents encounter problematic content, such as hate speech or conspiracy theories, regularly or repeatedly, suggesting that such exposure is embedded in everyday digital experience rather than sporadic. Behavioural science helps explain this pattern: people tend to stick with the options put in front of them, particularly when defaults and friction make it difficult to change settings. Without intervention, the path of least resistance is to passively consume what the algorithm serves up.

- *A narrow focus on fact-checking or the removal of illegal content addresses only a small part of the risk to civic discourse.*
- *Civic discourse on social media is being degraded* by the sheer volume of polarising commentary, hate speech, conspiracy theories, and hostility that remains within platform rules.
- *Repeated long-term exposure to such corrosive content* can shift perceptions of what is normal, and inflame and cement partisan divides, and thus have a significant negative effect on healthy civic discourse.
- *Engagement-optimised platform design could favour certain styles of political content.* The pattern is consistent with the idea that content optimised for outrage and 'us versus them' framing performs well in recommender systems. This dynamic is harmful to civic discourse.

2. Platform feeds are skewed towards right-wing content

The platform audit shows that young users' feeds are characterised by a disproportionate amplification of right-wing content. Bondata's survey complements this finding: 40 per cent of those strongly on the left said their feeds matched their views very poorly, compared with 8 per cent of those strongly on the right. From the perspective of civic discourse, exposure to viewpoints that differ from one's own is generally beneficial. However, when such exposure is concentrated at only one end of the political spectrum and not the other, it creates an unbalanced information environment.

A possible explanation is that right-wing actors may simply be more effective at using social media, demonstrating skills such as

being more adept with new technologies, and more persuasive in their messaging etc. However, the audit design challenges this explanation. The imbalance persisted even when avatars explicitly signalled interest in left-wing content, indicating that amplification patterns were not primarily driven by user behaviour. Moreover, this dominance appeared on average across platforms and country contexts. Yet it is not possible to conclude whether this reflects a real algorithmic bias, a skewed supply of content, hidden user preferences, or a mixture of different factors. Follow-up studies would be necessary to assess whether these results hold over time and to examine the underlying drivers in greater depth.

For civic discourse, the implications are, nevertheless, significant:

- *Pluralism may be undermined.* Democratic deliberation depends on balanced exposure to a wide range of viewpoints.
- *What constitutes 'mainstream' may be curated by an algorithm rather than people's views.* When one side of the political spectrum is systematically more visible, it may shift perceptions of what constitutes the 'mainstream.' Even without explicit persuasion, repeated exposure can normalise certain narratives and marginalise others. Over time, amplification itself may confer legitimacy.
- *Mitigation should target the system, not just content.* Potential bias – intentional or not – in recommender systems warrants further investigation, underscoring the importance of transparency and independent auditability.
- *Potential systemic risk under the EU Digital Services Act.* If platform feeds are structurally skewed towards any political orientation, this may constitute a systemic risk to civic discourse under the Digital Services Act.

3. Algorithmic unpredictability reduces user control and weakens agency

Algorithmic recommendations were highly unpredictable and opaque. Avatars' signals of interest did not consistently shape what appeared in feeds, and sometimes sudden shifts led to exposure to extreme content without any clear trigger. Bondata's survey data align with this pattern.

- *Algorithmic unpredictability undermines user agency.* If individuals cannot form a coherent understanding of how their behaviour shapes what they see, meaningful choice becomes difficult.
- *Lack of transparency in algorithms limits users' ability to curate their information environment and weakens trust in the neutrality of digital platforms.* From the perspective of democracy, transparency is not merely a technical requirement but a condition for informed participation.

4. Heavy social media use and emotional stress

Young people spend over 5 hours daily on social media platforms and encounter high volumes of problematic content. Survey responses suggest this environment is emotionally taxing. Around half of respondents to Bondata's survey reported negative emotions, such as disappointment, anger, fear, or sadness, when encountering political discussions on social media.

Considering this, it may initially appear contradictory that over 40 per cent of respondents nevertheless said that exposure to political content increases their willingness to participate in discussions. However, increased participation should not automatically be interpreted as a sign of healthy civic discourse. Anger, fear and disappointment may trigger reactive forms of participation – such as

commenting, sharing or confronting opposing views – which increase visible activity without necessarily supporting reflective or deliberative discourse.

- *If platforms profit from maximising engagement*, democratic discourse becomes collateral damage. This contributes to the degradation of the digital public sphere – sometimes referred to as 'enshittification'.
- *A harmful feedback loop could be set in motion.* Emotionally charged content drives interaction; interaction becomes a ranking signal, the system amplifies similar content, and users encounter more of it, creating a self-reinforcing cycle that can intensify polarisation over time.
- *Negative emotions can translate to disengagement.* A very large share of young social media users reported negative feelings about the political discussions on social media – and some report reduced willingness to participate. Platform dynamics are plausibly depressing constructive civic engagement and trust.

5. AI-generated political content lowers the threshold for manipulation

Of all political posts, 5 per cent were clearly AI-generated, with an additional 2 per cent suspected. These included deepfakes of politicians, AI-generated meme videos, and synthetic avatars expressing hostile or extremist views. The presence of AI-generated content was particularly worrying in the second audit, when most of such content was labelled as hostile speech, misinformation, or simply had a malicious message. Although still small, this category is significant and the amount of AI-generated content is growing due to the proliferation of technology.

- *Generative AI lowers the cost* of producing persuasive, emotionally charged and highly shareable content at scale. It also enables rapid variation, i.e. multiple versions of the same message tailored to different audiences – making influence operations cheaper, faster, and harder to detect.
- *As detection becomes more challenging*, the distinction between authentic and synthetic political expression may blur further.
- *The interaction between the Digital Services Act and the Artificial Intelligence Act* will be crucial in addressing this emerging risk, particularly where AI-generated content may contribute to systemic distortions of civic discourse.

6. A degraded social media environment increases external influence risks

The current social media environment does not only harm domestic democratic debate – it also creates structural vulnerabilities for external actors to shape perceptions, norms and political outcomes. The proliferation of AI further amplifies these risks.

This vulnerability is not limited to traditionally adversarial actors such as Russia, which has repeatedly used disinformation and online manipulation to destabilise European democracies. It also extends to shifts in geopolitical alignment more broadly. For example, the recently published United States National Security Strategy (The White House 2025) emphasises strategic competition in the information domain and expresses optimism about "growing influence of patriotic European parties". Recognising these risks, the European Union has launched initiatives such as the Democracy Shield to strengthen democratic resilience against external interference and systemic risks to civic discourse.

- *Influence operations become cheaper and more effective, and harder to attribute* when platforms amplify extreme content, normalise hostility and fuel negative emotions. These dynamics can be used to foster cynicism towards democratic institutions.
- *Young people are a high-value target for external influence operations.* Social media is the primary arena in which young people encounter political information, form attitudes and engage in public discussion.

7. Structural risks require structural responses

Taken together, the findings suggest that risks to civic discourse arise less from isolated falsehoods or illegal content, and more from systemic features of how platforms are built and run.

The harm lies not only in misinformation, but in how content is selected, framed, and amplified – and in how largely hidden recommender systems interact with predictable human tendencies to follow the path of least resistance. These dynamics are closely linked to engagement-driven business models, where revenue depends on maximising attention, time spend, and data collection.

- *The primary risks to democracy and civic discourse are structural.* The most consequential harms stem from platform design and functioning: ranking, amplification, defaults, and incentives, more than from individual illegal posts.
- *User agency must be made real.* Transparent systems, meaningful controls, safer defaults and investments in digital literacy are essential, especially for young users.
- *Democratic oversight of platforms is essential to safeguarding democracy.* Regulators, supervisory authorities, journalists, and researchers must be able to scrutinise – and, where necessary, help reshape—the social media platforms that now function as a central arena for civic discourse.

Concluding remarks

In early February 2026, the European Commission announced its preliminary finding that TikTok may be in breach of the Digital Services Act due to features considered 'addictive by design', including infinite scroll, autoplay, push notifications, and a highly personalised recommender system. The European Commission stated in its risk assessment that the company 'did not adequately assess how these addictive features could harm the physical and mental wellbeing of its users, including minors and vulnerable adults' (European Commission 2026).

The Commission's announcement is a significant step towards demonstrating that systemic design choices – not only individual posts – fall within the scope of democratic oversight. History repeats itself, and in some ways social media can be compared to tobacco: first an extremely addictive product is created, harms are downplayed, research findings are withheld, and policy-makers are lobbied to keep them from intervening. Now the tide is beginning to turn, as the harms of social media have become too visible to ignore.

Ultimately, the task for democratically elected policy-makers is to reclaim oversight of the democratic process, including in the digital sphere. This requires political institutions to assert their authority, enforce existing rules rigorously and, where necessary, redesign the regulatory framework so that power over the digital public sphere is no longer in the hands of a few digital giants. Without decisive action, the cumulative effects of imbalance, lack of transparency and emotional degradation risk reshaping civic culture in ways that democratic institutions may later struggle to correct.

5. Recommendations

Based on this study, seven recommendations can be drawn.

1. Enforce DSA transparency and user-control requirements

The EU should continue to rigorously enforce the Digital Services Act to ensure that very large online platforms provide clear, user-friendly explanations of how their recommender systems work, as well as equip users with meaningful tools to have influence over what they see. Further action is needed. In line with the requirements of the DSA, platforms should disclose the main ranking parameters in plain language, offer adjustable settings, and provide a non-profiling feed option. Transparency should go beyond formal compliance through standardised reporting on, for example, amplification patterns.

2. Ensure independent, long-term systemic risk auditing

Systemic risks to civic discourse require continuous, independent monitoring. The EU should fully enforce Article 40 of the DSA by ensuring vetted researchers and regulators have effective access to very large online platforms data for research related to systemic risks. In particular, sustained monitoring is needed to track ideological amplification – including potential political bias – exposure to problematic and AI-generated content, and longer-term emotional and behavioural impacts.

3. Reform online choice architectures to strengthen user agency

The EU should require very large online platforms (VLOPs) to adopt protective defaults – such as reduced autoplay and notification intensity, clear and easily accessible content controls, and simple tools to adjust recommendation settings. Platforms should also remove 'sludge' that makes it hard to change defaults. Behavioural insights should inform DSA risk-mitigation so that design supports informed choice rather than exploiting cognitive biases for engagement.

4. Strengthen democratic resilience through digital information literacy, epistemic rights and civic tech

Digital literacy must evolve beyond technical skills. While young people are agile users of digital services, educational systems and civil society organisations need to equip them with new capabilities and critical thinking for online environments. Promoting young Europeans' democratic engagement through participatory democracy, including so called civic tech platforms, is important. Individuals need agency to be active citizens with a sense of ownership of their democracy. Mechanisms such as citizen panels, citizen initiatives and online deliberation can strengthen that agency, and thereby help citizens in taking back control of the civic discourse. The EU Democracy Shield initiative underscores the role of

citizen engagement and participation in protecting and renewing democracy. The EU should integrate epistemic rights[4] into digital governance frameworks, ensuring citizens have access to truthful information and the ability to understand how AI systems affecting public life are developed and used.

5. Coordinate enforcement of the DSA and the AI Act to address AI-generated political content

As AI-generated political content becomes more prevalent, it can lower the threshold for manipulation and intensify amplification, exacerbating the systemic risks covered by the DSA. The EU and Member States should coordinate DSA and AI Act enforcement through strong cross-border cooperation, require clear labelling and traceability of AI-generated political content, and build the technical capacity to assess recommender systems, audit algorithmic risks, and verify platform compliance.

6. Reduce structural lock-in and support pluralistic digital ecosystems

To curb systemic risks driven by market concentration, the EU should strengthen user mobility and digital self-determination by expanding data portability beyond personal data, developing privacy-preserving standards for optional reputation portability, and explicitly recognising protection from manipulative design as a democratic right. EU institutions and Member States should also diversify their official communications across alternative platforms. Reducing lock-in lowers dependence on digital giants and strengthens democratic sovereignty over the digital public sphere.

7. Protect minors through enforceable, privacy-preserving age assurance

Given evidence of emotional strain and susceptibility to manipulative design, the EU and Member States should consider raising and effectively enforcing minimum age limits for full-feature social media access, preferably through coordinated action at the EU-level. Platforms should be mandated to implement privacy-preserving age-verification systems (such as trusted third-party assurance using tokenised verification) and publish transparency reports on enforcement accuracy. The objective is to shield young users from exploitative design – not to restrict legitimate democratic participation.

4 Epistemic rights refer to the requirement that to have equality in decision-making, our societies should guarantee that truthful information and knowledge are made available to all its citizens and that they have the competence to use these for their own benefit and that of society as a whole (Nieminen, 2024).

References

Arora, S. D., Singh, G. P., Chakraborty, A., & Maity, M. (2022). Polarization and social media: A systematic review and research agenda. Technological Forecasting and Social Change, 183, 121942. https://www.sciencedirect.com/science/article/abs/pii/S0040162522004632

Augenstein, I., Bakker, M., Chakraborty, T., Corney, D., Ferrara, E., Gurevych, I., Hale, S., Hovy, E., Ji, H., Larraz, I., Menczer, F., Nakov, P., Papotti, P., Sahnan, D., Warren, G., Zagni, G. (2025). Community Moderation and the New Epistemology of Fact Checking on Social Media. Cornell University. Article 26 May 2025.

Barberá, P. (2020). Social media, echo chambers, and political polarization. Social media and democracy: The state of the field, prospects for reform, 34-55. https://www.cambridge.org/core/books/social-media-and-democracy/social-media-echo-chambers-and-political-polarization/333A5B4DE1B67EFF7876261118CCFE19

Berger, J., & Milkman, K. L. (2012). What makes online content viral?. Journal of Marketing Research, 49(2), 192-205.

BIT. (2019). Improving consumer understanding of contractual terms and privacy policies: evidence-based actions for businesses. https://www.bi.team/wp-content/uploads/2019/07/BIT_WEBCOMMERCE_GUIDE_DIGITAL.pdf

BIT. (2024a). Are content controls the answer to helping people curate their online feeds? https://www.bi.team/blogs/are-content-controls-the-answer-to-helping-people-curate-their-online-feeds/

BIT. (2024b). More than half the public think they're good at spotting false information online, but only a third think other people are. BI Team Press Release. https://www.bi.team/press-releases/more-than-half-the-public-think-theyre-good-at-spotting-false-information-online-but-only-a-third-think-other-people-are/

BIT. (2024c). The shrouded economy. https://www.bi.team/wp-content/uploads/2024/03/Shrouded-Economy-Working-Paper.pdf

BIT. (2024d). Terms & Conditions Apply. https://www.bi.team/blogs/terms-conditions-apply/

BIT. (2025). Behavioural Audit of online services. https://www.ofcom.org.uk/online-safety/safety-technology/behavioural-audit-of-online-services

Bradford, A. (2023). Digital Empires. The Global Battle to Regulate Technology. Oxford University Press

Bursztyn, L., Handel, B. R., Jimenez, R., & Roth, C. (2023). When product markets become collective traps: The case of social media (No. w31771). National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w31771/w31771.pdf

Cheng, Z., Chen, J., Peng, R. X., & Shoenberger, H. (2023). Social media influencers talk about politics: Investigating the role of source factors and PSR in Gen-Z followers' perceived information quality, receptivity and sharing intention. Journal of Information Technology & Politics, 21(2), 117-131. https://www.researchgate.net/profile/Zicheng-Cheng/publication/368233745

Claud, F. (2022). Effects of Social Media Algorithms on the Political Perspectives of 4th Year Political Science Students. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4839376

Doctorow, Cory (2025). Enshittification: Why Everything Suddenly Got Worse and What to Do About It. 7.10.2025. MCD.

Dufva, M., Kiiski-Kataja E., Lähdemäki-Pekkinen J. Megatrends 2026. 28.1.2026. Sitra studies 253.

Drolsbach, C., & Pröllochs, N. (2025). Characterizing AI-Generated Misinformation on Social Media. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2505.10266

European Commission. (2025a). European Democracy Shield: Empowering Strong and Resilient Democracies. JOIN (2025) 791 final.

European Commission. (2025b). Youth survey 2024. https://europa.eu/eurobarometer/surveys/detail/3392

European Commission. (2026). Commission preliminarily finds TikTok's addictive design in breach of the Digital Services Act. 6.2.2026.

Eurostat. (2025). Young people - digital world. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Young_people_-_digital_world#:~:text=By%202024%2C%20the%20gap%20between,65%25%20of%20the%20total%20population.

Flayelle, M., Brevers, D., King, D. L., Maurage, P., Perales, J. C., & Billieux, J. (2023). A taxonomy of technology design features that promote potentially addictive online behaviours. Nature Reviews Psychology, 2(3), 136-150. https://www.researchgate.net/profile/Maeva-Flayelle/publication/368510187_A_taxonomy_of_technology_design_features_that_promote_potentially_addictive_online_behaviours/links/63eca87b19130a1a4a7d9379/A-taxonomy-of-technology-design-features-that-promote-potentially-addictive-online-behaviours.pdf

Global Witness. (2025). X and TikTok algorithms push pro-AfD content to non-partisan German users. Global Witness Press Release. https://globalwitness.org/en/press-releases/x-and-tiktok-algorithms-push-pro-afd-content-to-non-partisan-german-users-new-analysis/

Goujard, C., Braun, E., & Scott, M. (2024). TikTok and the far right in European Parliament politics. Politico Europe. https://www.politico.eu/article/tiktok-far-right-european-parliament-politics-europe/

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., ... & Tucker, J. A. (2023a). How do social media feed algorithms affect attitudes and behavior in an election campaign?. Science, 381(6656), 398-404. https://www.science.org/doi/10.1126/science.abp9364

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., ... & Tucker, J. A. (2023b). Reshares on social media amplify political news but do not detectably affect beliefs or opinions. Science, 381(6656), 404-408. https://www.science.org/doi/10.1126/science.add8424

Huertas, Jaime Ferrando (2023). X's open source algorithm - Unveiling the code, but not the secrets. 31.3.2023. https://www.shaped.ai/blog/twitters-open-source-algorithm-unveiling-the-code-but-not-the-secrets. (retrieved on 24th February 2026)

Hughes, Laura, Borress, Amy (2024). Teenage social media use strongly linked to anxiety and depression. Financial Times. 13.10.2024.

Härkönen, T., Lehtonen, K. Vahti, J., Vänskä, R. (2022). Tracking Digipower – How data can be used for influencing decision-makers and steering the world. 23.5.2022 Sitra studies 215.

Ibrahim, H., Jang, H. D., Aldahoul, N., Kaufman, A. R., Rahwan, T., & Zaki, Y. (2025). TikTok's recommendations skewed towards Republican content during the 2024 US presidential race. https://arxiv.org/abs/2501.17831

Karlsen, R., & Aalberg, T. (2023). Social media and trust in news: An experimental study of the effect of Facebook on news story credibility. Digital Journalism, 11(1), 144-160. https://www.tandfonline.com/doi/full/10.1080/21670811.2021.1945938

Kubin, E., & von Sikorski, C. (2021). The role of (social) media in political polarization: a systematic review. Annals of the International Communication Association, 45, 188-206. https://www.tandfonline.com/doi/full/10.1080/23808985.2021.1976070

Laibson, D. (1997). Golden eggs and hyperbolic discounting. The Quarterly Journal of Economics, 112(2), 443-478. https://dash.harvard.edu/server/api/core/bitstreams/7312037c-7431-6bd4-e053-0100007fdf3b/content

Levy, R. E. (2020). Social media, news consumption, and polarization: Evidence from a field experiment. American Economic Review, 111(3), 831-870. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3653388

Loewenstein, G., Sunstein, C. R., & Golman, R. (2013). Disclosure: Psychology changes everything. 18.8.2013 Harvard Public Law Working Paper No. 13-30. https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=2312708

Mackinac Center. (N.d.) The Overton Window. https://www.mackinac.org/OvertonWindow#resources (retrieved on 6 October 2025)

Madriaza, P., Hassan, G., Brouillette-Alarie, S., Mounchingam, A. N., Durocher-Corfa, L., Borokhovski, E., ... & Paillé, S. (2025). Exposure to hate in online and traditional media: A systematic review and meta-analysis of the impact of this exposure on individuals and communities. Campbell systematic reviews, 21(1), e70018. https://onlinelibrary.wiley.com/doi/pdf/10.1002/cl2.70018

Marquart, F., Ohme, J., & Möller, J. (2020). Following politicians on social media: Effects for political information, peer communication, and youth engagement. Media and Communication, 8(2), 197-207. https://www.cogitatiopress.com/mediaandcommunication/article/view/2764

Mosseri, A. (2021). Shedding more light on how Instagram works. Instagram Blog. (retrieved on 20 June 2025) https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works

Muth, L., & Peter, C. (2023). Social media influencers' role in shaping political opinions and actions of young audiences. Media and Communication, 11(3), 164-174. https://www.cogitatiopress.com/mediaandcommunication/article/view/6750

Mäkelä, R-M., Tähkäpää, O., Vahti, J. Foresight review: Transformation of the security environment. 8.4.2026. Sitra Foresight Review #3.

Nieminen, H. (2024). Why we need epistemic rights? in Horowitz, M., Nieminen, H., Lehtisaari K., D'Arma, A. (Ed.) Epistemic Rights in the Era of Digital Disruption, Palgrave MacMillan. https://library.oapen.org/handle/20.500.12657/86930

Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., ... & Tucker, J. A. (2023). Like-minded sources on Facebook are prevalent but not polarizing. Nature, 620(7972), 137-144. https://pubmed.ncbi.nlm.nih.gov/37500978/

Oden, A., Porter, L. (2023). The kids are online: Teen social media use, civic engagement, and affective polarization. Social Media + Society, 1-12. https://journals.sagepub.com/doi/10.1177/20563051231186364

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. Nature Communications, 13(1), 233/3. https://www.nature.com/articles/s41467-022-30073-5.pdf

Quickframe. (2025). The Twitter algorithm: Understanding how it works. Quickframe Blog. https://quickframe.com/blog/the-twitter-algorithm/

Rekola, S., Tuori, S., Vahti, J. Future barometer 2025 (Tulevaisuusbarometri 2025). 12.3.2025. (Accessible only in Finnish.) Sitra publication 246.

Reuters. (2026). US to fund free speech initiatives in Europe, Trump official says. 9.2.2026 (retrieved on 18 February 2026). https://www.reuters.com/world/us-fund-free-speech-initiatives-europe-trump-official-says-2026-02-09/

Schmuck, D., Hirsch, M., Stevic, A., & Matthes, J. (2022). Politics—simply explained? How influencers affect youth's perceived simplification of politics, political cynicism, and political interest. The International Journal of Press/Politics, 27(3), 738-762. https://journals.sagepub.com/doi/10.1177/19401612221088987

Smith, B. (2021, December 5). How TikTok Reads Your Mind. The New York Times. https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html

Sultan, M., Tump, A. N., Ehmann, N., Lorenz-Spreen, P., Hertwig, R., Gollwitzer, A., & Kurvers, R. H. (2024). Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors. Proceedings of the National Academy of Sciences, 121(47). https://www.pnas.org/doi/10.1073/pnas.2409329121

Swart, J. (2021). Experiencing algorithms: How young people understand, feel about, and engage with algorithmic news selection on social media. Social Media + Society, 7(2). https://www.researchgate.net/publication/350827693_Experiencing_Algorithms_How_Young_People_Understand_Feel_About_and_Engage_With_Algorithmic_News_Selection_on_Social_Media

Swart, J. (2023). Tactics of news literacy: How young people access, evaluate, and engage with news on social media. New media & Society, 25(3), 505-521. https://journals.sagepub.com/doi/10.1177/14614448211011447

Terren, L., & Borge, R. (2021). Echo chambers on social media: A systematic review of the literature. https://www.researchgate.net/publication/349945771_Echo_Chambers_on_Social_Media_a_Systematic_Review_of_the_Literature

Thaler, R. H., & Shefrin, H. M. (1981). An economic theory of self-control. Journal of Political Economy, 89(2), 392-406.

Thaler, R. H. (2018). Nudge, not sludge. Science, 361(6401), 431-431. https://www.science.org/doi/full/10.1126/science.aau9241

Thaler, R. H., & Sunstein, C. R. (2021). Nudge: The final edition. Penguin.

TUI Stiftung (2025). Jugendstudie 2025 Junge Menschen: EU und Demokratie sind gut, es braucht aber Reformen. 3.7.2025. (accessed on 24th February 2026)

V-Dem Institute (2025). Democracy report 2025. March 2025.

Wasastjerna, M. (2020). Competition, Data and Privacy in the Digital Economy. Wolters Kluwer.

World Health Organization (2024). A focus on adolescent social media use and gaming in Europe, central Asia and Canada: Health Behaviour in School-aged Children international report from the 2021/2022 survey. Vol. 6. WHO Regional Office for Europe.

The White House (2025). National Security Strategy of the United States. November 2025. https://www.whitehouse.gov/wp-content/uploads/2025/12/2025-National-Security-Strategy.pdf

Ye, J., Luceri, L., & Ferrara, E. (2024). Auditing Political Exposure Bias: Algorithmic Amplification on Twitter/X Approaching the 2024 US Presidential Election. https://arxiv.org/abs/2411.01852

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. Journal of personality and social psychology, 9(2p2), 1. https://web.mit.edu/curhan/www/docs/Articles/biases/9_J_Personality_Social_Psychology_1_%28Zajonc%29.pdf

Zinigrad, L. (2024, September 2). How social media amplifies support for the far right in France. LSE European Politics and Policy Blog. https://blogs.lse.ac.uk/europpblog/2024/09/02/how-social-media-amplifies-support-for-the-far-right-in-france/

Annex – platform audit methodology

Designing the audit

This annex describes a summary of the audit methodology used by BIT, based on a longer Technical Appendix. The Technical Appendix in its entirety is published separately.

Instagram, Tiktok, and X were selected for the audit based on their popularity among EU youth, and/or for their influence on news dissemination, political discourse, and real-time commentary, ensuring variety in ownership and content type. The following algorithmically curated feeds were audited.

- Instagram: Reels
- TikTok: For You page
- X: Main feed

The audit was carried out in Finland, France, and Romania. These countries were chosen with a view to provide a geographical, cultural and political diverse sample.

BIT created two avatars aged 18–24 for each platform in each country, totalling 24 avatars.

Avatar engagement

At sign-up, avatars selected 'news' as their interest and followed three well-known and established news sources from their countries widely considered as 'independent' from political movements. The news outlets followed are listed in the Technical Appendix.

All avatars followed some general rules of engagement, guided by the research aims as well as BIT's safeguarding and ethical principles. Although these guidelines were standardised across countries, readers should note that cross-country comparisons are influenced by differing political contexts and by researcher variation, which may lead to subtle differences in platform interaction.

BIT conducted seven sessions with each account varying how the avatars engaged with political content.

- In the 'Low-Engagement' phase, avatars did not show any particular interest in politics.
- In the 'High-Engagement' phase, avatars followed major political parties across the spectrum and watched political posts for longer than other types of content.
- In the 'Tilted Trajectory' phase, avatars maintained a higher watch time for political content, and one avatar went on a left-wing trajectory, the other on a right-wing trajectory. These meant that the avatar signalled interest exclusively in either left-of-centre or right-of-centre views or topics by unfollowing some of the political parties. In Romania, instead of a left–right split, we used a pro-EU and EU-sceptic distinction. During the Tilted Trajectory phase, the pro-EU avatar unfollowed EU-sceptic parties, while the EU-sceptic avatar unfollowed pro-EU parties. This reflected the Romanian context, where attitudes towards the EU provide a more meaningful distinction than the traditional left–right divide.

The audit was conducted in two distinct phases in 2025:

- First audit: Conducted during the Summer across all three countries.
- Second audit: Conducted in Finland leading up to Christmas.

In both audits, avatars went through the 'Low-Engagement', 'High-Engagement', and 'Tilted Trajectory' phases, as detailed above.

The second audit aimed to verify the robustness of previous findings in Finland by introducing three methodological changes. First, avatars followed a new left-of-centre party previously excluded from the audit. Second, researchers reclassified another party from 'centrist' to 'left-of-centre'. Third, the avatars liked the five most recent posts of the parties they followed in the 'Tilted Engagement' phase — something they did not do in the first audit. This sent a stronger signal about what political content they are interested in. Further details about these methodological changes can be found in the Technical Appendix. The following sections indicate whether specific findings apply to both audits or are unique to one.

Categorisation of political content

We documented and categorised the political content avatars saw during each audit journey. Key information captured included:

- Whether the post could be attached to right-wing, left-wing or centrist politics;
- Whether the post included e.g. mis- or malinformation, conspiracy theory, or hostile or hate speech.

Table 5. Some key definitions of terms used in this publication.

Term	Definition
AI-generated content	Content that was either labelled as created or edited using artificial intelligence on the platform, or content that researchers judged to be clearly created or edited using artificial intelligence.
Conspiracy theory*	A belief or explanation that attributes the cause of an event or situation to the secret and coordinated actions of powerful actors – such as governments, corporations, or elites – typically without solid evidence and often in contradiction to the mainstream or official account.
Extremist content	Content that spreads or indicates support for views outside the political mainstream, including: • Openly anti-democratic views; • Racism and hatred against groups based on protected characteristics; • Support for radical far-left or far-right movements, parties or governments (historical or present), such as Nazism and communism. Note that such content does not necessarily violate the guidelines of platforms, especially if such views are expressed in covert, indirect ways.
Hate speech*	Public incitement to violence or hatred directed against a group of persons or a member of such a group defined on the basis of protected characteristics, such as gender, race, colour, descent, religion or belief, or national or ethnic origin.
Hostile speech*	Content that expresses malevolence, discriminates against or mocks a group of persons or a member of such a group defined on the basis of protected characteristics.
Left-wing, centrist, or right-wing content	The categorisation of political content into these buckets was based on: • Widely used definitions, such as those in the Encyclopedia Britannica; • Local contexts, including political affiliations of the content creator or the use of arguments or language associated with a particular political movement in the given country.
Malinformation*	Information that is based on a fact but is presented in a misleading way, e.g. removed from its context, in order to purport a specific narrative or perspective.
Misinformation*	Content including or based on false information.
Political content*	Content that mentions governments, elections, or social topics (e.g. immigration, reproductive rights, climate change, LGBT rights, inequality, racism, taxes).
Problematic content	Content that researchers categorised as falling into one or more of the following categories: AI-generated content, conspiracy theory, extremist content, hate speech, hostile speech, malinformation, misinformation. We use this term freely to refer to content that might pose risk to civic discourse, might violate platform guidelines, or might be upsetting, misleading or disturbing to users.

**Note: These definitions were developed to balance and reflect definitions provided by EU institutions, the platform's community guidelines, as well as the expert judgement of BIT's researchers specialised in this field.*

Analysis

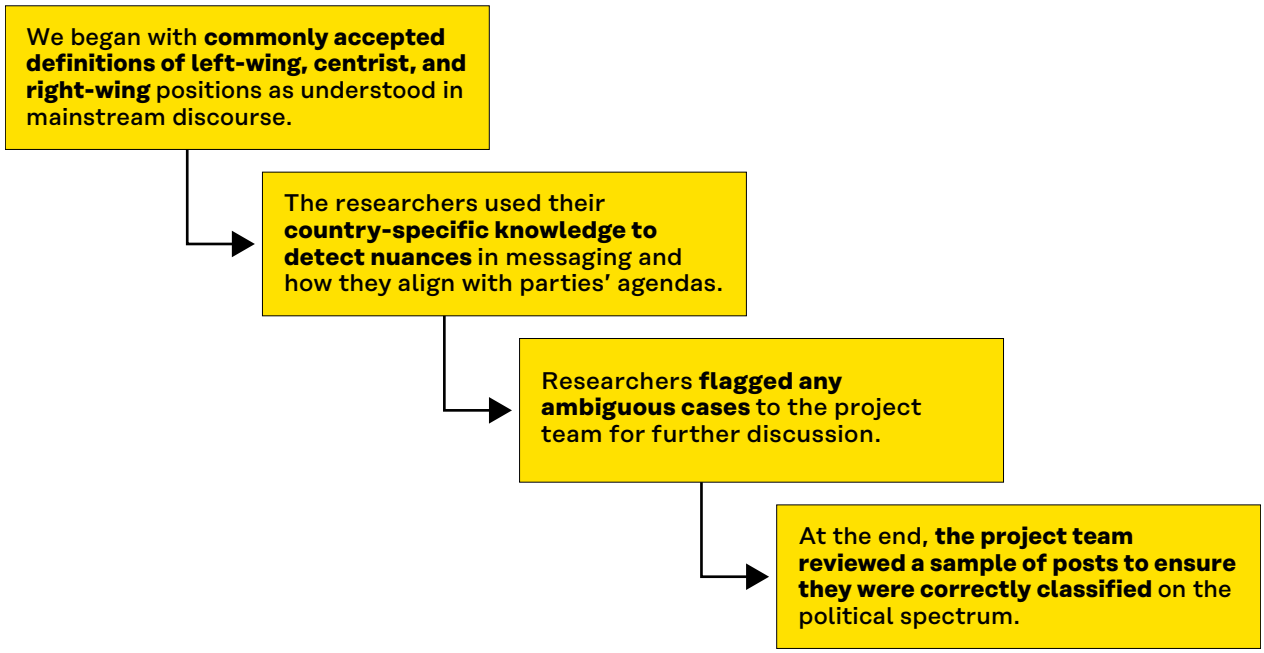
At the end of the first data collection period, BIT arrived at a dataset containing 1,337 pieces of political content (also referred to as 'political posts' throughout the report), collected by the 18 avatars across 126 browsing sessions. The dataset contained detailed information about each political post, including a description of its content and various categorisations.

BIT analysed this dataset through a combination of qualitative and quantitative analysis. To perform qualitative analysis, we grouped the political posts by the browsing session they appeared in. We manually analysed the content, language use, and inferable or stated political ideology of each of the posts. Data from different countries were coded by different researchers. At the end of the coding process, we compared the categorisations of posts to ensure consistency across researchers and made changes to the coding as appropriate. Figure below outlines the exact process of categorising political content along the political spectrum.

Then, BIT conducted quantitative analysis to gain further insights. This involved producing descriptive statistics (e.g. the number of political posts containing right-wing, centrist and left-wing narratives) and using AI and machine learning techniques to assign themes to each post. A small subset of the posts could not be labelled automatically and were assigned a theme manually.

We applied the same techniques to analyse the dataset from the second audit. Results and interpretations were updated based on the additional insights gained from this analysis.

Figure 9. The process of categorising content along the political spectrum.



More detailed descriptions of the avatar journeys on each platform and in each country, together with selected screenshots, are provided in a separate Technical Appendix.

In addition to the platform audit, Bondata, a Finnish research firm, conducted an online survey on young adults' social media use. The survey covered the same countries as the BIT audit but used a broader age range (18–29) to obtain more representative samples. In total, 3,063 responses were collected: Finland (1,030), France (1,022), and Romania (1,011). At the 95% confidence level, the margin of error for the full sample is ±3.1 percentage points. BIT's platform audit and Bondata's survey represent very different research methods. While their results are presented and analysed together, the reader should keep in mind these methodological differences.

About the Behavioural Insights Team authors

Sujatha Krishnan-Barman is a Principal Advisor and the Head of Consumers and Business Markets at BIT. She leads the organisation's work in digital markets, business productivity, and consumer protection. She is particularly interested in online choice architecture and crosscountry regulatory approaches to deal with novel challenges. She has previously worked as a behavioural insights consultant, a macroeconomist covering South Asia, and an investment banker. Prior to joining BIT, Suze completed her doctorate in cognitive neuroscience at University College London where her research centered on exploring naturalistic social interactions involving both neurotypical adults as well as those with autism spectrum condition. She also holds an MSc in Cognitive and Decision Sciences from UCL, an MSc in Comparative Politics from the LSE as well as a degree in management from the Indian Institute of Management in Ahmedabad.

Bálint Dercsényi is an Advisor based in the London office, working in the Consumer & Business Markets team and the Gambling Policy & Research Unit. He has experience across a range of policy areas, including online safety, gambling, and business productivity. Bálint holds an MSc in Behavioural Science from the London School of Economics and an MA (Hons) in Economics and German from the University of St Andrews.

Laurence Fenn is an Associate Advisor based in the London office, working in the Consumer & Business Markets team across a range of policy areas. His recent work includes delivering a behavioural audit of popular social media platforms for the UK media regulator, Ofcom. He has also conducted behavioural audits under BIT's Gambling Policy Research Unit, examining online instant win games and safer gambling tools using an innovative methodology to assess user journeys and consumer risk. Prior to joining BIT, Laurence worked as an auditor at a Big Four accountancy firm in Cambridge. He holds a First Class Honours in BSc Management from the London School of Economics and a Distinction in MSc Economics from the University of St Andrews.

Cindia Li was formerly an Associate Advisor in BIT's Consumer and Business Markets team.

BIT is a global research and innovation consultancy which combines a deep understanding of human behaviour with evidence-led problem solving to improve people's lives.

SITRA

SITRA STUDIES 256
The Sitra studies series publishes the results of Sitra's future-oriented work and experiments.

ISBN 978-952-347-457-4 (PDF) www.sitra.fi

SITRA.FI
Itämerenkatu 11–13
PO Box 160
00181 Helsinki
Tel: +358 294 618 991
(X handle) @SitraFund