Algorithms and democracy: How social media shapes young Europeans' worldviews

Technical Appendix

Authors: Bálint Dercsényi, Laurence Fenn, Sujatha Krishnan-Barman and Cindia Li

Contents

Appendix 1: Evidence review and Methodology

Evidence review

BIT conducted a rapid review of the available evidence on how young people consume news and political information, the kind of content served to social media users and how this might shape civic discourse. We reviewed academic publications, high-quality investigative journalism, and fact-checked reports from reputable outlets. Table 1 summarises the key findings from this review and the quality of evidence available on each focus area.

Table 1. Summary of key findings from the evidence review and the quality of evidence

Topic	Key finding	Quality of evidence	Notes on quality of evidence
Young people's sources of political content	Social media has become the main source of political content for young people across the EU	High	Clear findings from an EU-survey
Content served to young people on social media	The political content shown on social media is often extremist, more often right-wing than centrist or left-wing, and often misleading	Medium	Findings across several sources point to a similar direction, but most sources are informal (e.g. journalistic investigation) and often not specific to young people or the EU
Consumption of news on social media	Consumption of political news on social media is characterised by disengagement, scepticism and lack of trust.	High	Aligning findings across qualitative and quantitative studies in the EU and other European countries
The overall effect of social media use on polarisation	Studies show a mixed picture. Some studies show that social media use contributes to certain types of polarisation, but others find no effects, or in some cases, even reduced polarisation. It is unclear under what circumstances social media polarises.	Medium	There is a relatively large body of evidence on this topic, including RCTs and literature reviews. But the overall impact remains unclear. Studies use different methodologies, measure different types of polarisation, and test different platforms in different countries,

Topic	Key finding	Quality of evidence	Notes on quality of evidence
			making comparison difficult.
Mechanisms used by platforms to decide the type of content served	All three platforms' recommendations are driven by engagement-based signals, e.g. watching and liking content. However, not enough is known for independent researchers to reproduce the full user journeys. Out of the three platforms of interest, the least is known about X's mechanisms.	Medium	Given limited platform transparency, most evidence comes from journalistic investigation.
Platform-specific partisan bias in political content exposure	TikTok and Twitter exhibit severe biases towards far-right content. This appears to be less pronounced with Instagram.	Medium	The available evidence is limited to a few countries and in some cases not peer-reviewed (e.g. journalistic evidence).
Mechanisms that determine whether social media increases or decreases polarisation	Two main possible mechanisms identified: echo chambers and exposure to counter-attitudinal views. However, it's still contested how widely spread these phenomena are and how they influence polarisation or political views.	Medium	There is a relatively large body of evidence on this topic, including RCTs. But the overall impacts remain unclear. Studies use different methodologies, measure different types of polarisation, and test different platforms in different countries, making comparison difficult.
The role of influencers in shaping political views	The findings indicate that influencers shape young people's views because they deliver simple political messages, and users have a positive psychological relationship with them and their content (e.g. see them as credible).	Medium	The reviewed studies have similar findings about why influencers are particularly well-placed to influence young people's attitudes. However, much of the evidence is survey- or interview-based, with limited experimental evidence.
Specific algorithmic changes	Existing testing of specific changes to algorithmic design (e.g. removing reshared content, introducing chronological feeds, or reducing exposure to like-minded	Medium	Evidence is mostly limited to studies testing specific interventions in a single country on a single platform.

Topic	Key finding	Quality of evidence	Notes on quality of evidence
	material) have shown limited success in curbing polarisation or exposure to misinformation.		
Regulatory approaches	Existing regulatory approaches are specific to single jurisdictions and their effectiveness have not been evaluated. Suggested strategies centre around greater user control, greater platform transparency, content labelling, and coordinated international efforts.	Low	No formal evaluations of implemented policies. Current discourse is mainly theoretical/ based on expert opinion.
Media literacy initiatives	Existing evidence suggests that media literacy initiatives are effective in improving critical assessment of online news. However, literature also critiques that these initiatives are inadequate on their own as a mitigation strategy, as they place the onus on individuals, and tend to be particularly ineffective for the politically disengaged majority who may be unaware of filter bubble effects.	Medium	Robust evidence exists on the effectiveness of media literacy initiatives on improving news discernment in itself, but very little is known about whether these initiatives are sufficient as a mitigation strategy in the context of real-life social media use and its broader impact on civic discourse.

Methodology

The following sections provide details on how we planned and executed the audit of social media platforms.

We selected Instagram, Tiktok, and X for the audit based on their popularity among EU youth, ensuring variety in ownership and content type. We audited the following algorithmically curated feeds:

- 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. The audit was conducted by researchers fluent in the local language and familiar with the political and cultural environments of the respective countries, ensuring accurate interpretation and coding of social media content.

General approach

We conducted the audit by creating new avatar accounts, two for each platform in each country, totalling 18 avatars. Researchers signed up to the platforms and conducted seven separate sessions with each account over the course of 3-4 weeks. Across the sessions, researchers varied how the avatars engaged with the content on the platforms. In the Audit journeys section we describe the types of sessions completed.

After the completion of the first audit, Sitra commissioned BIT to conduct a second audit in Finland only. The aim of this second audit was to verify the robustness of previous findings in Finland by introducing some methodological changes. These changes are detailed in the Audit journeys section.

Sign-up

Researchers created accounts either by using SIM cards purchased for this research for verification or, when that wasn't possible, by signing up with new email accounts created for this research.

Avatar specifications

The avatars were 18-24 years old (date of birth was randomly allocated within this range). None of the audited platforms required information about gender at sign-up, so this information was not provided. The exact avatar specifications are outlined in Table 2 below.

Table 2. Avatar specifications

Country	Date of birth	Age	Political leaning	Avatar name	Avatar username
First Audit					
Finland	14/03/2003	22	left	Lumi Korhonen	l.korhonen101
Finland	09/11/2004	20	right	Mika Virtanen	m.virtanen102
Romania	26/11/2001	23	left	Alex Popa	a.popa103
Romania	22/05/2007	18	right	Gabi Dumitru	g.dumitru104
France	30/09/2005	19	left	Camille Lemoine	c.lemoine105
France	02/01/2006	19	right	Sacha Moreau	s.moreau106
Second Audit					
Finland	16/02/2002	23	left	Aino Nieminen	A.Nieminen101
Finland	02/09/2005	20	right	Onni Laine	O.Laine101

Researchers did not provide any data to platforms beyond what was required for account creation. They rejected optional data sharing options, cookies, and used an incognito mode browser.

Account setup

X required new users to indicate interest in at least one topic. All avatars selected 'news' as their interest, which aligned with our research focus, but did not predestine the avatars to see certain types of political content on their feeds.

All new avatars followed three well-known news sources from their countries independent from any political movements. The aim of this was to:

1. Satisfy X's requirements to follow at least one account and keep this consistent across platforms;
2. Signal a general interest in current affairs, without aligning with political parties, movements, world views, or specific political issues.

The avatars followed the following outlets:

- Finland: YLE Uutiset, Helsingin Sanomat, MTV Uutiset
- France: Le Monde, Agence France Presse, France Info
- Romania: ProTV, TVR1, DIGI24

Technical specifications

- **Devices**: BIT staff used company laptops with incognito browsers; one subcontractor used a personal device with incognito browsers. TikTok was audited using mobile apps, as detailed later in this section.
- **VPN usage**: Researchers not located in target countries used VPNs to connect to servers in their assigned country; this was the case for Romania and Finland.
- **Screen recording**: All sessions and sign-ups were screen-recorded with audio for subsequent coding.
- **Language settings**: Platform languages were set to local languages on Instagram and TikTok, and to local languages plus English on X, where this was possible.

General rules of engagement

Avatars followed some general rules of engagement (see Table 3), guided by the research aims as well as our safeguarding and ethical principles described below.

Avatars first completed any setup-related actions, such as following the selected news accounts. Afterwards, they browsed the selected feeds on each platform.

They were not allowed to use any other platform functionalities, such as direct messaging or searching for hashtags.

While browsing, avatars watched content and occasionally opened comment sections (see details in the Audit journeys section for how this differed across the types of journeys). They did not comment on or reshare posts, and only liked a small set of posts in the second audit, which had been shared by party accounts. Engagement with content aimed to simulate normal user behaviour, including continuous scrolling and lingering on videos for longer than picture- or text-based content, but was strictly limited to actions needed for the research's purposes.

Table 3. Rules of engagement

Type of engagement	Rules of Engagement
Sign-up	All avatars must be set to **private** if the platform allows. This reduces the visibility of the avatar, and potential for other users to interact with the avatar.

Following accounts	Avatars should only follow accounts as set out below:
	• At **sign-up**, follow three neutral news channels / public service broadcasters with a large following:
	  ○ Finland: Yle, MTV and Helsingin Sanomat
	  ○ France: Le Monde, France Info, AFP
	  ○ Romania: ProTV, TVR1, DIGI24
	For audit 1
	• When you enter the **high-engagement phase** (before session 3), follow the official accounts of the political parties listed below for your country.
	  ○ Finland: Green Party, SDP, Suomen Keskusta, Kokoomus, Perussuomalaiset
	  ○ France: LFI, EN, RN
	  ○ Romania: USR, PNL, AUR, PSD

- When you enter the **tilted trajectory phase** (before session 5), unfollow parties as shown below to signal the avatar's interest:
    - Finland
        - Left-of-centre interest (e.g. environmental protection, minority rights, social support): unfollow Suomen Keskusta, SDP, Kokoomus, Perussuomalaiset.
        - Right-of-centre interest (e.g. business-friendly policies, stronger borders): unfollow the Green Party, Suomen Keskusta, and SDP.
    - France
        - Left-of-centre interest: unfollow EN and RN.
        - Right-of-centre interest: unfollow LFI and EN.
    - Romania
        - Pro-EU interest: unfollow AUR and PSD.
        - EU-sceptic interest: unfollow USR, PNL, and PSD.

For Audit 2

- When you enter the **high-engagement phase** (before session 3), follow the official accounts of the political parties listed below for your country.
    - Finland: Green Party, Vasemmistoliitto, SDP, Suomen Keskusta, Kokoomus, Perussuomalaiset
- When you enter the **tilted trajectory phase** (before session 5), unfollow parties as shown below to signal the avatar's interest:
    - Finland
        - Left-of-centre interest (e.g. environmental protection, minority rights,

	social support): unfollow Suomen Keskusta, Kokoomus, Perussuomalaiset. ■ Right-of-centre interest (e.g. business-friendly policies, stronger borders): unfollow the Green Party, Vasemmistoliitto, SDP and Suomen Keskusta. At this stage, like the five most recent posts by the party accounts still followed. Note that not all parties had an account on all the audited platforms. When there was not an official party account to follow, the avatars followed the account of the party leader instead.
Engagement with children online	The avatar should not follow any accounts that appear to be owned by someone under the age of 25. When assessing whether an account holder is likely to be under 25, researchers will take into account indicators such as references to recent birthdays, mentions of school or university activities, handles that suggest birth years or graduation dates. In cases where the avatars receive direct contact (e.g. direct messages) from accounts that likely belong to children that indicates risks of harm, researchers should refer to the instructions outlined in the row below on 'direct messages'.
Engaging with content	The avatars must not like, comment on, or reshare any content. The only exception to this is liking posts shared by followed political parties in the second audit (see the Audit journeys section for details).

	Engagement will come from viewing content and following accounts.
Reporting content	Researchers may encounter harmful or illegal content. Should this occur, researchers must **follow the risk-appropriate safeguarding actions outlined in the safeguarding protocol.**
Direct messages	Avatars must not initiate direct messages with other users or respond to direct messages received from other users. However, if avatars receive direct messages that indicate that **a real user is at risk of harm**, researchers should follow the steps below: • Do NOT respond to the message through the avatar account • Screenshot the interaction • Assess the risk level against the safeguarding protocol and take the appropriate associated actions • If the account is potentially owned by a child (i.e. someone who appears to be under 25), take extra care to assess the content and err on the side of over-escalation

Audit journeys

Each avatar journey consisted of seven browsing sessions, grouped into three phases. Browsing sessions lasted 10 minutes or until seeing 20 political posts, whichever happened first. A small number of sessions slightly deviated from this length due to technical difficulties or human error.

The avatars changed their behaviours and signalled interests across these phases to mimic a new user gradually developing interest in politics. Below we describe each stage.

- In the 'Low-Engagement' phase (two sessions), avatars did not show any particular interest in politics.
- In the 'High-Engagement' phase (two sessions), avatars followed major political parties across the spectrum and watched political posts for longer than other types of content.
- In the 'Tilted Trajectory' phase (three sessions), 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.
    - In the second audit, avatars sent an even stronger signal by not just following, but also liking the five most recent posts of political parties aligning with their interest.

See Table 3 for more details on which parties were followed at each stage.

Coding

Researchers used codebooks to document and categorise the political content they saw during each audit journey. Filling out the codebooks involved recording high-level information about the journey, such as the number of political and non-political posts seen, as well as a detailed description and categorisation of each political post. Researchers recorded:

- What the post was about, including country-specific context;
- Who it was posted by (account handle);
- Whether the post's main message or the opinion expressed could be attached to right-wing, left-wing, or centrist politics;
- Whether the post included misinformation, conspiracy theory, or hate speech (see the Annex of the main report for the definitions used);
- Any comments justifying the categorisation or raising concerns and uncertainties.

Ethical considerations and safeguarding principles

BIT is committed to conducting research ethically and to the highest standard. This project was subject to our research ethics process, which meets the criteria set out by the UK Government's Social Research (GSR) Unit[1], the Market Research Society (MRS) Code of Conduct[2] and the Economic and Social Research Council's (ESRC) guidance on governance arrangements for research ethics committees.[3] Our ethics policies are regularly updated ensuring alignment with GSR. We have a research ethics panel consisting of trained staff who conduct project reviews. To ensure the independence of the ethics panel, only panel members who are not involved in the project in question can review it.

We planned and conducted the audit in accordance with the following two main principles:

1. Minimising the impact of the research on any real users;
2. Avoiding any negative impacts on the researchers conducting the audit.

To minimise the research's impact on real users, we adopted strict rules of engagement (outlined in Table 3), which restricted what avatars were and were not allowed to do on the platforms. In essence, we avoided any unnecessary interaction with real users and avoided amplifying any content that might be illegal, go against platforms' community guidelines, or constitute misinformation, hate speech, or conspiracy theory. Therefore, researchers were advised not to like, comment on, or reshare any posts and only follow accounts belonging to mainstream media outlets or major political parties. Note that the French avatars on TikTok needed to like some content in order to signal interest in certain types of content, as they could not follow accounts on the platform.

We have also developed a safeguarding protocol (see Table 4 below), giving instructions to researchers distressed by the content seen on the platforms or encountering content that requires escalation. Researchers were advised to stop auditing if feeling distressed and reach out to either the project lead or to BIT's

[1] *Ethical Assurance Guidance for Social Research in government*. (2011). GOV.UK. Available here.
[2] MRS Code of Conduct (Mat 2023). Available here.
[3] *Governance arrangements for research ethics committees* (n.d.). UK Research and Innovation. Available here.

designated safeguarding lead, who could offer appropriate next steps. Researchers were also advised to report any content clearly violating community guidelines, following the platforms' standard procedures, then resume auditing. Finally, researchers were directed to further guidance in case they encountered content that required immediate attention to help prevent serious harm, such as credible threats of self-harm, content related to human trafficking, child exploitation, or terrorist threats.

Table 4. Safeguarding protocol

Risk	Safeguarding actions
Researchers encounter **low-severity harmful content or distressing content** online such as: • Fake news or misinformation • Potential scam or phishing attempt • Exposure to content that may negatively shape body image (e.g. heavily edited or filtered content) • Indications of loneliness, stress, low mood	No formal safeguarding actions are required. If distressed, you are encouraged to: • Take a short break from data collection • Raise it with the project lead for support or to discuss adjusting your role Document relevant examples at the end of data collection to support broader analysis.
Researchers encounter **moderate-severity harmful content** online that violates platform-specific community guidelines, but are not illegal, such as: • Exposure to hateful content online • Exposure to adult pornography • Indicators of bullying, harassment, or non-acute mental health distress • Posts suggesting substance abuse or emotional distress	If content is reportable under platform rules, file a report using standard platform procedures. If the content originates from a user who may be a child, in a vulnerable situation, indicates that they might be at risk of harm, or is messaging directly, notify the project lead. Include screenshots and brief context. You may pause your session and check in with the team if you feel affected.

• AI-generated misleading content, e.g. impersonation of public figures without declaration that the content is AI-generated Under EU law, illegal content refer to terrorist content, xenophobic or racist hate speech, child sexual abuse, breaches of intellectual property rights, and unsafe products (Eucrim, 2018)	Returning to browsing should only happen if and when you feel comfortable.
Researchers encounter **high-severity harmful or illegal content** online such as: • Terrorist content • Child abuse material • Extreme and child pornography • Explicit references to self-harm, suicide, or ongoing abuse • Indicators of human trafficking or imminent danger (especially involving children)	Immediately stop browsing and notify the project lead and safeguarding lead via secure channels. Do not download or forward illegal content. Instead, document enough detail (e.g. platform, username, post timestamp) to enable follow-up. Report the content through platform procedures. Escalate to external agencies as needed: • For researchers in the UK, please refer to Ofcom's published guidance. • For researchers in the EU, please refer to Europol guidance for country-specific links to report. The safeguarding lead will advise whether legal consultation or external reporting is required. Researchers may pause or opt out of continuing with this session entirely.

Appendix 2: Findings from individual journeys

The following sections illustrate the browsing experience in each journey, across the three countries and three platforms. The descriptions of the journeys are complemented by screenshots of notable posts seen by the avatars and charts showing the most prominent themes in political posts in each browsing session.

Trigger warning:

Some of the screenshots in this appendix display mis- and malinformation, conspiracy theories, and hate speech or hostile speech. Others use strong language or make offensive, insensitive jokes concerning topics such as the Holocaust, religious, sexual and ethnic minorities.

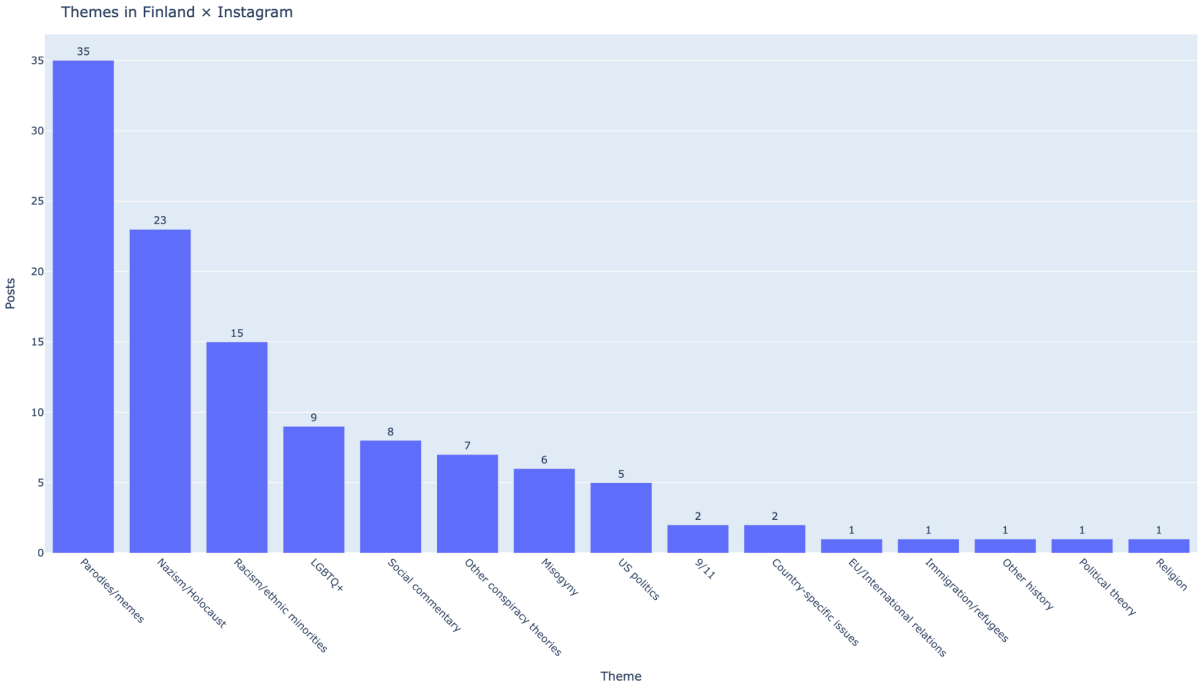
Audit 1

Finland

Instagram

The vast majority of political content encountered by the Finnish avatars on Instagram was humour-based extreme right-wing content. This included both memes or expressions of views using meme formats, and covered topics such as Nazism and The Holocaust, racism, sexual minorities and women (see Figure 8 for the thematic breakdown of posts). These posts also frequently included hate speech and hostile speech, misinformation, and conspiracy theories (see the descriptions of individual journeys below for details). Both the right- and left-wing avatars saw the same types of posts, indicating that expressing interest in left- versus right-wing politics had no effects on the content shown in the Reels feed. There was one notable difference between the two journeys: while the left-wing avatar saw this kind of problematic content from the very first browsing session, the right-wing avatar saw no political content in its first six sessions before the extremists memes appeared for the seventh and final session.

Figure 8: Thematic analysis of all political posts (Finland, Instagram)



Right-wing avatar

This user journey (see Figures 9 and 10) contained no political content up until the last browsing session. In this session, without any obvious trigger, there was a sudden and dramatic change in the recommended content. In this session, political posts dominated the feed. The majority of these posts were memes or posts using meme formats, referencing Nazism, hatred against various minority groups, and the Second World War. These posts were very similar to the ones shown to the other Finnish avatar on Instagram. For example, posts joked about bullying transgender people and indicated that being racist or a Nazi supporter is funny and or a positive thing. It remains unclear what triggered the algorithm to recommend this type of content in the last browsing session.

Figure 9: User journey illustration (Finland, Instagram, Right-wing avatar)

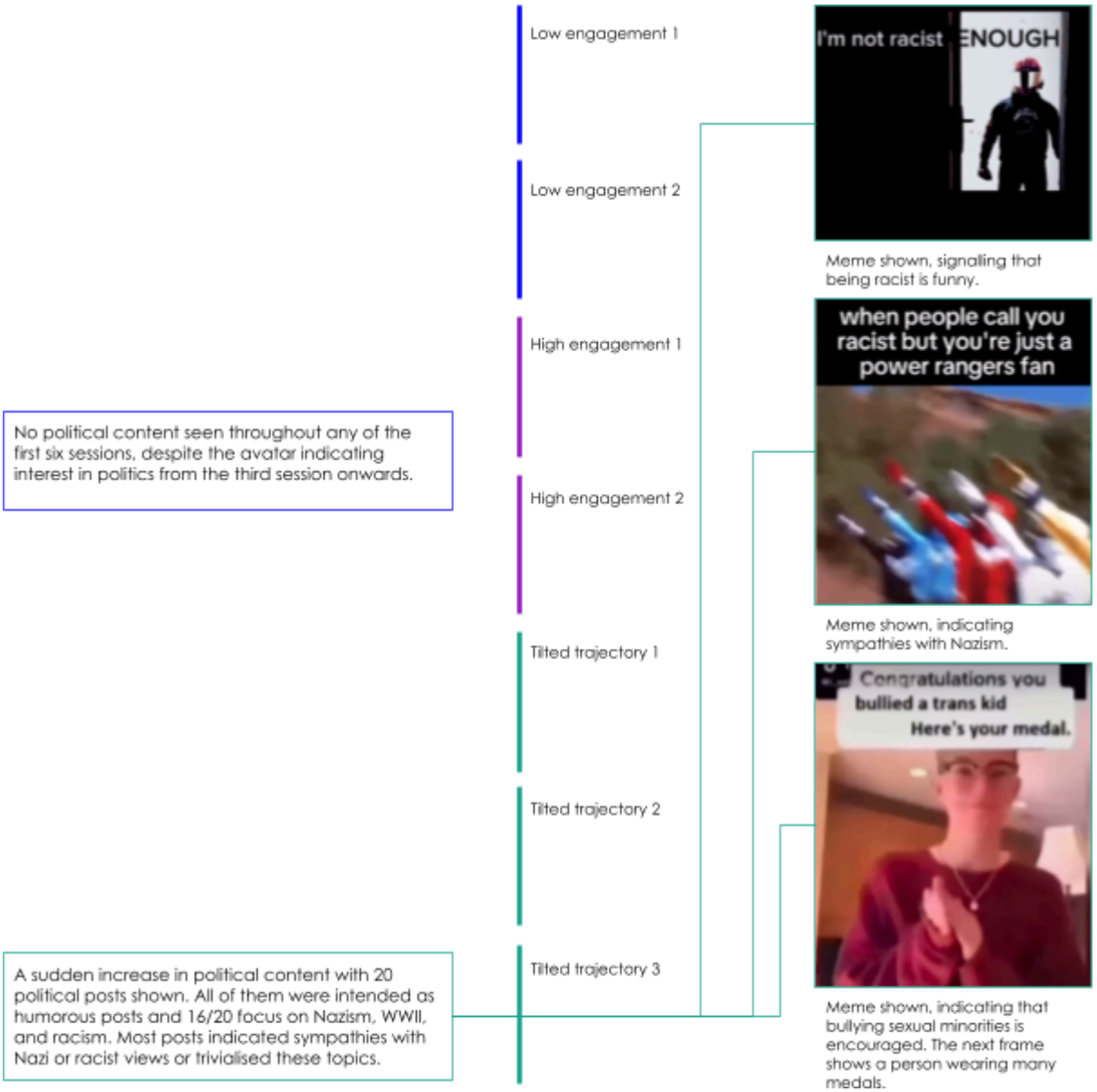
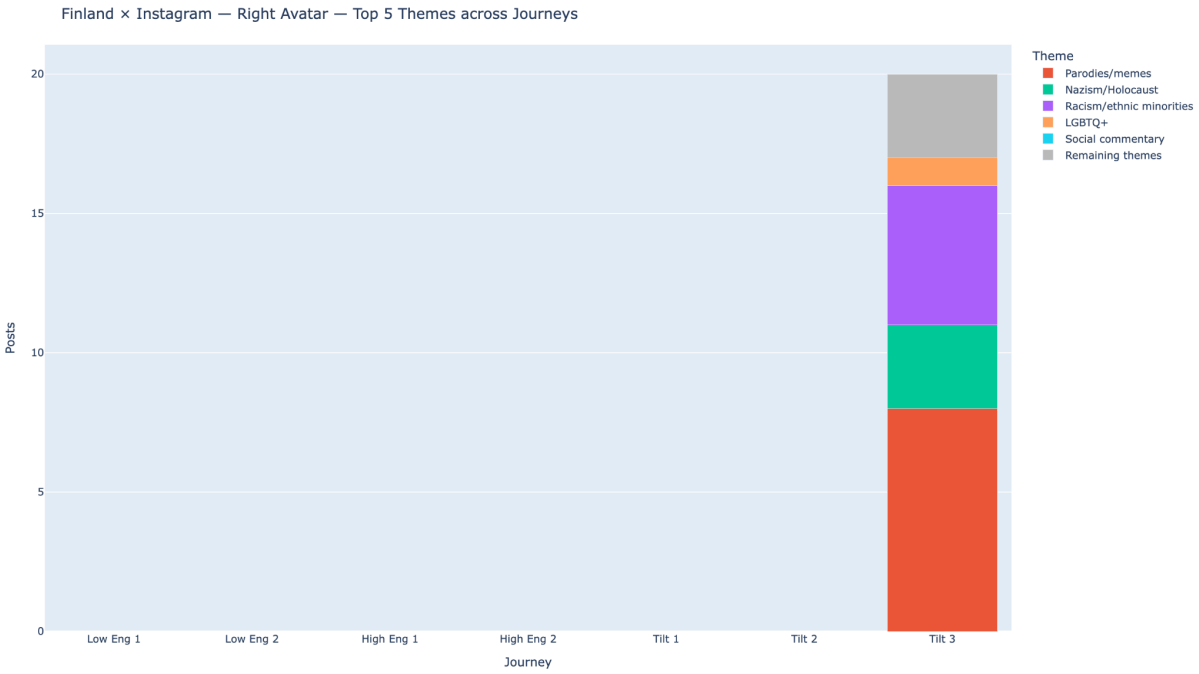


Figure 10: Thematic analysis of political posts encountered in user journey (Finland, Instagram, Right-wing avatar)



Left-wing avatar

This user journey (see Figures 11 and 12) contained by far the most extreme and problematic content. Throughout this journey, the amount of political content in the feed increased gradually. Initially, the avatar only saw three political posts in a ten-minute browsing session, while the final sessions were completely dominated by political content. Many were memes or serious posts using meme formats, differentiating them from more traditional political arguments made on other platforms. The content became increasingly radical. The initial journeys contained potentially offensive jokes and dark humour, but not hate speech, for example, a video of creating an 'unwearable T-shirt' labelled 'Adidolf' and showing Nazi salutes in the place of Adidas stripes. However, as the journey progressed, there were more and more posts expressing explicit support for Nazism. Posts containing outright hate speech emerged in the final session too. Most were directed against Jewish people and Black people, but multiple posts were hostile towards women, members of the LGBTQ+ community, and Muslims too. The avatar's stated interest in left-wing politics had no influence on the recommended content. In fact, in the 'tilted trajectory'

sessions, the avatar saw nothing but extremist right-wing content, without ever signalling such preferences. The posts also contained some conspiracy theories and misinformation, for example, a post that made fun of and grossly misrepresented the number of Holocaust victims murdered in gas chambers in the Second World War.

Figure 11: User journey illustration (Finland, Instagram, Left-wing avatar)

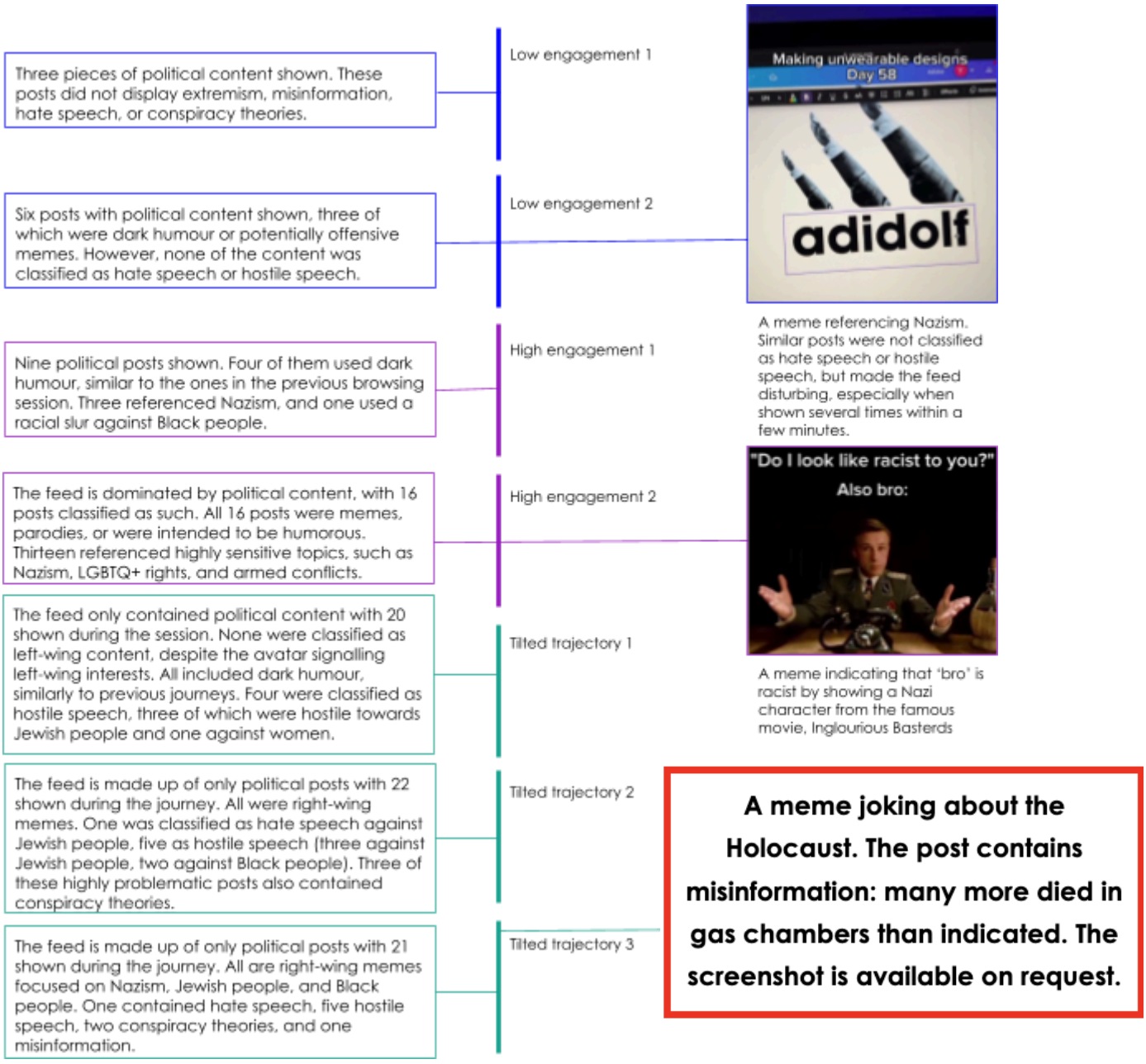
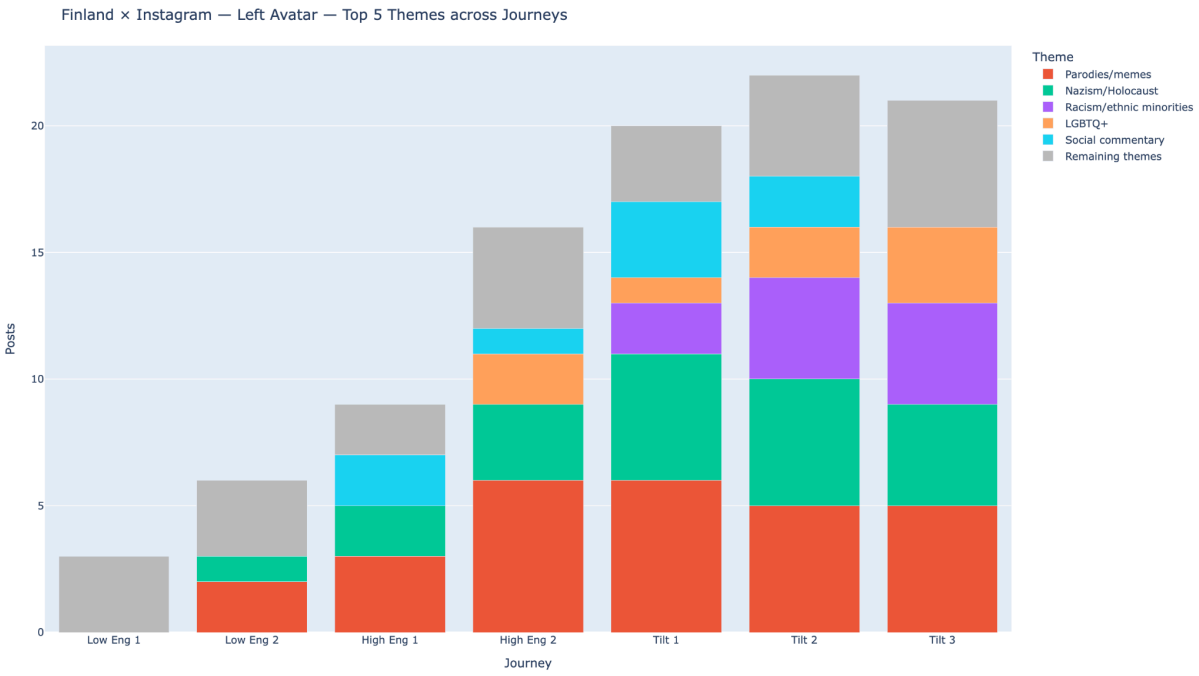


Figure 12: Thematic analysis of political posts encountered in user journey (Finland, Instagram, Left-wing avatar)

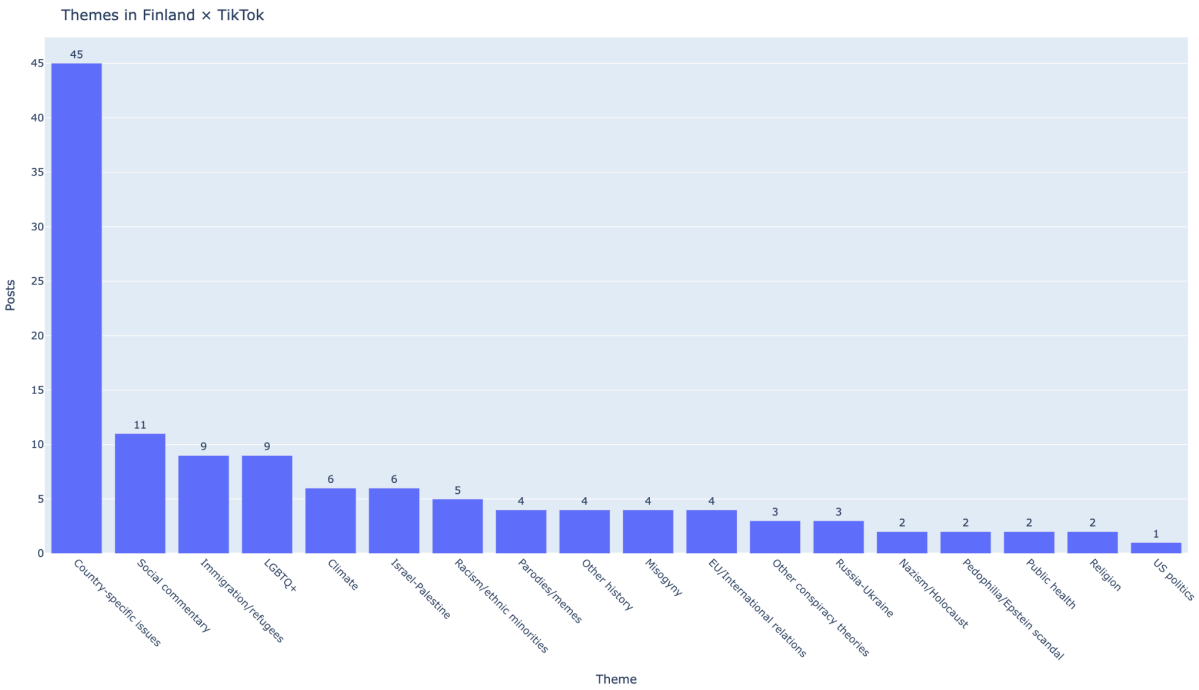


TikTok

The volume of political content on TikTok fluctuated heavily across browsing sessions. Both avatars observed a sudden spike in the number of political posts in their second session. This was still a 'Low Engagement' session, meaning that the avatars had not indicated interest in politics when the spike happened. After this, each session contained some political content, albeit at more moderate levels.

The content shown occasionally included problematic materials, such as deep fake videos of politicians and hateful posts towards ethnic minorities. However, there were no clear patterns in the types of problematic content displayed. Overall, the journeys included posts with both right- and left-wing views, roughly with the same frequency. Most of the posts (45 of 122) covered country-specific issues (see Figure 13).

Figure 13: Thematic analysis of all political posts (Finland, TikTok)

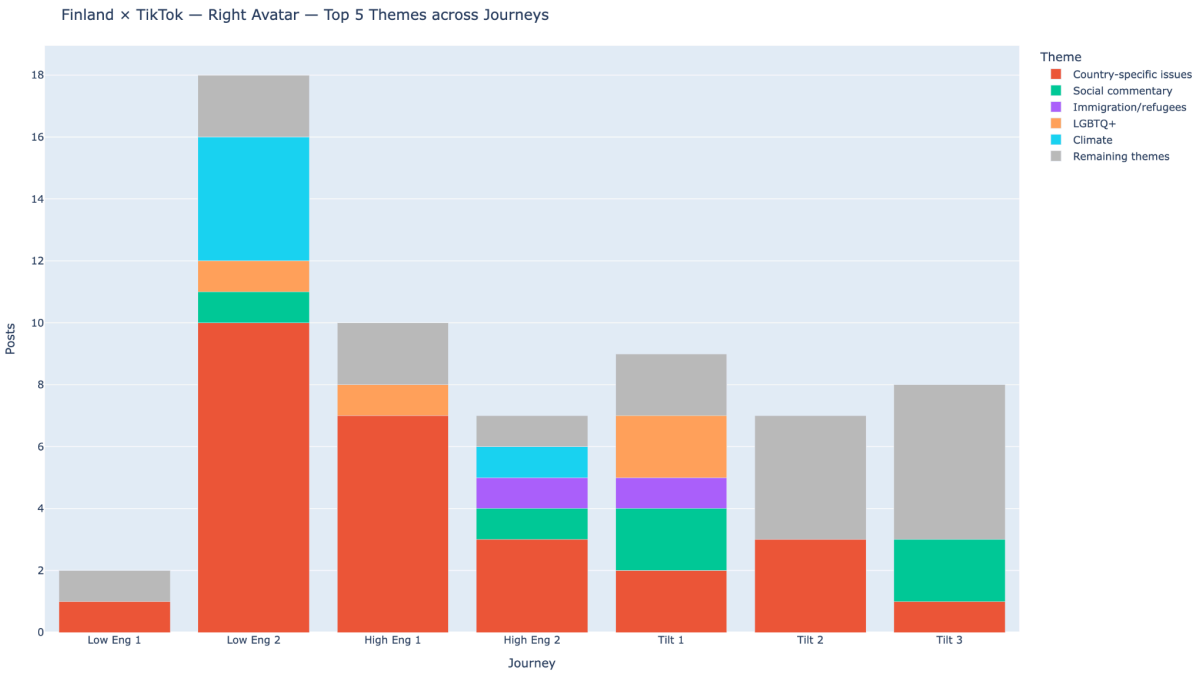


Right-wing avatar

At the start of this user journey (see Figures 14 and 15), the feed contained very little political content, although there was an AI-generated deep fake video of a former Finnish president giving a speech. The second browsing session brought a sudden spike in political content, more than half of which discussed issues specific to Finland. Contrary to the wider trends in this research, most of the posts in this session aligned with left-wing politics. Across the remaining sessions, the amount of political content dropped to a moderate level, covering a diverse range of topics — albeit local politics remained widely discussed. Overall, both left-wing, centrist, and right-wing views were represented, but this was spread unevenly across sessions. For example, the last session was dominated by left-wing content, while the session preceding it had none. The avatar's stated interest in general politics did not increase the amount of political content recommended, and later, the avatar's signalled interest in right-wing politics had no impact on the amount of right-wing content shown.

Figure 14: User journey illustration (Finland, TikTok, Right-wing avatar)

Figure 15: Thematic analysis of political posts encountered in user journey (Finland, TikTok, Right-wing avatar)



Left-wing avatar

This user journey (see Figures 16 and 17) was also characterised by sudden changes in the amount of political content shown, similarly to the TikTok journey with the other, right-leaning Finnish avatar. The first session contained few political posts, one of which was the same AI-generated deep fake that was shown to the other avatar as well. The second session brought a sudden increase in the amount of political content shown, much of which was extremist or categorised as hate speech or hostile speech. For example, one post joked about hatred against Black people, using a slur. The remaining sessions had fewer political content and fewer problematic posts. Initially, right-wing posts were more prominent, but once the avatar expressed interest in left-wing politics, the sessions became more balanced. This change was atypical: unfollowing parties did not lead to detectable changes in most other journeys. The last session contained only one political post, despite continued interest in politics by the avatar. This was another AI-generated post, showing a fake speech by the former Finnish president, Urho Kekkonen.

Figure 16: User journey illustration (Finland, TikTok, Left-wing avatar)

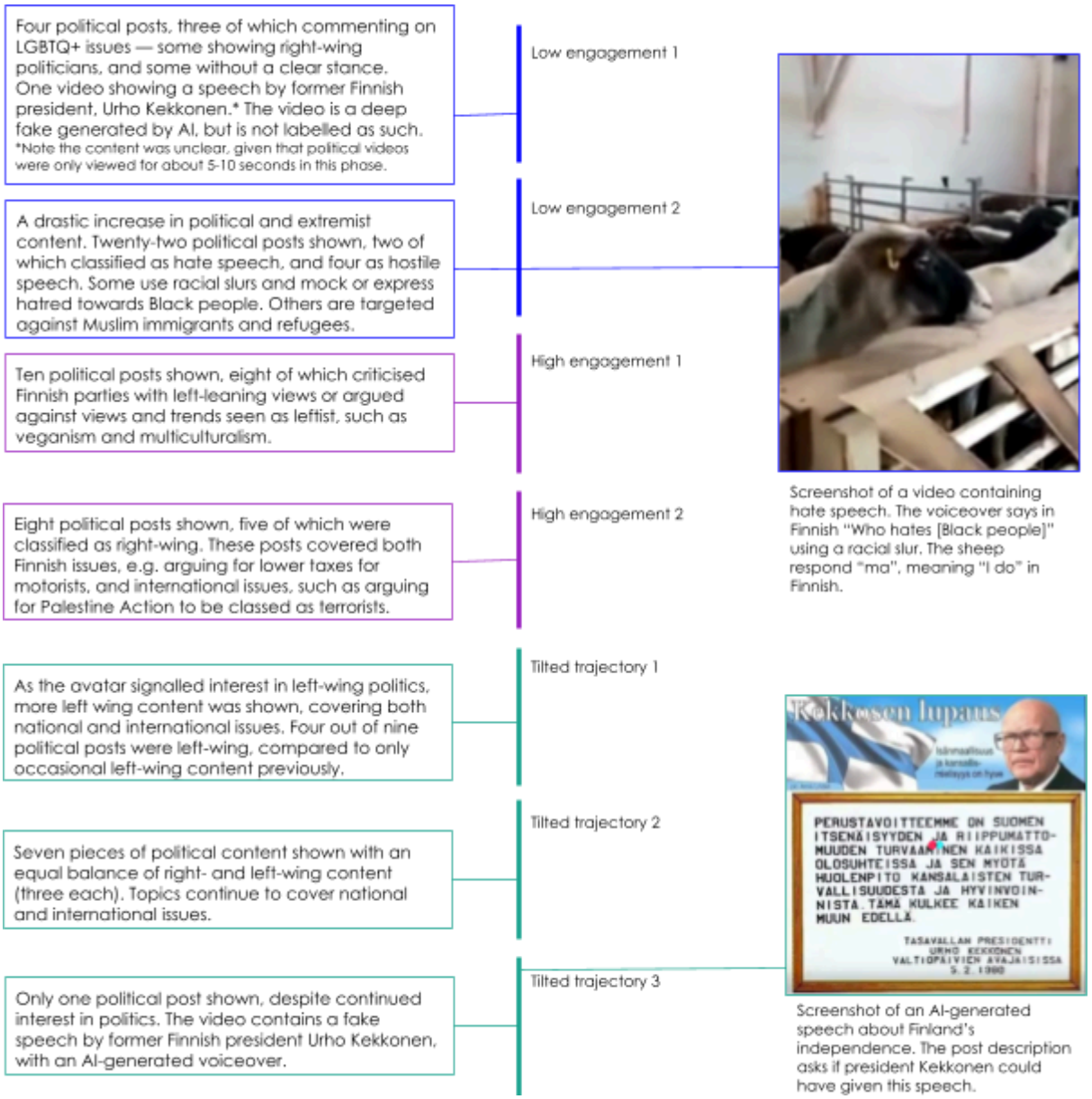
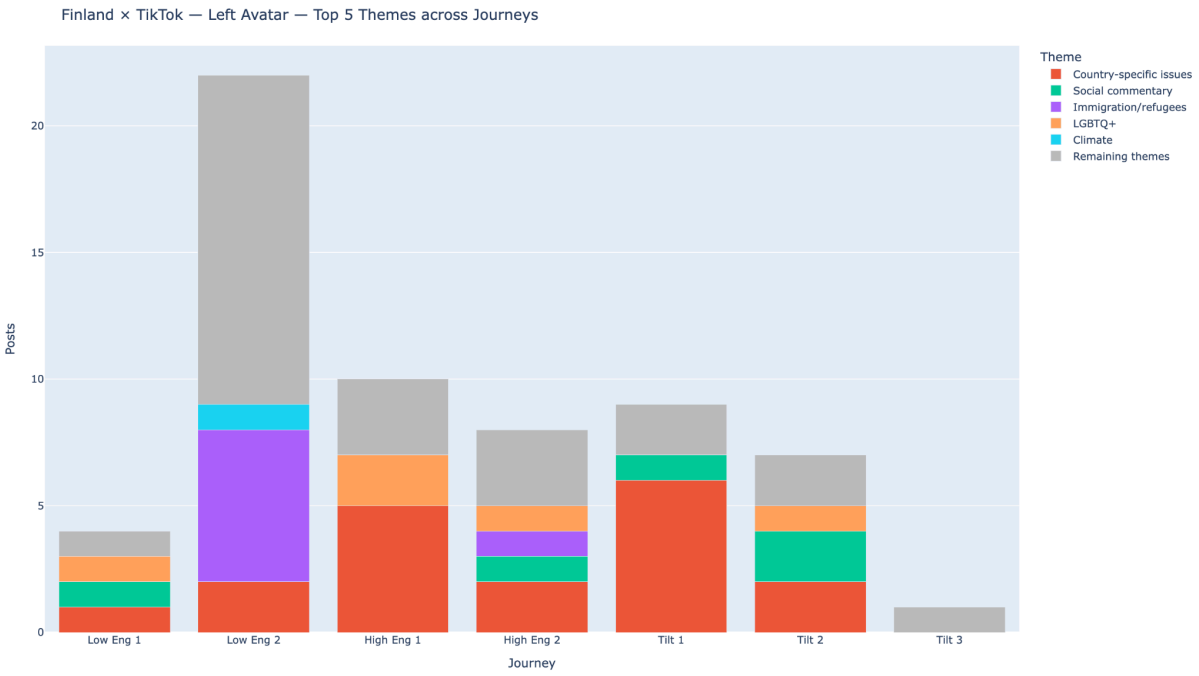


Figure 17: Thematic analysis of political posts encountered in user journey (Finland, TikTok, Left-wing avatar)

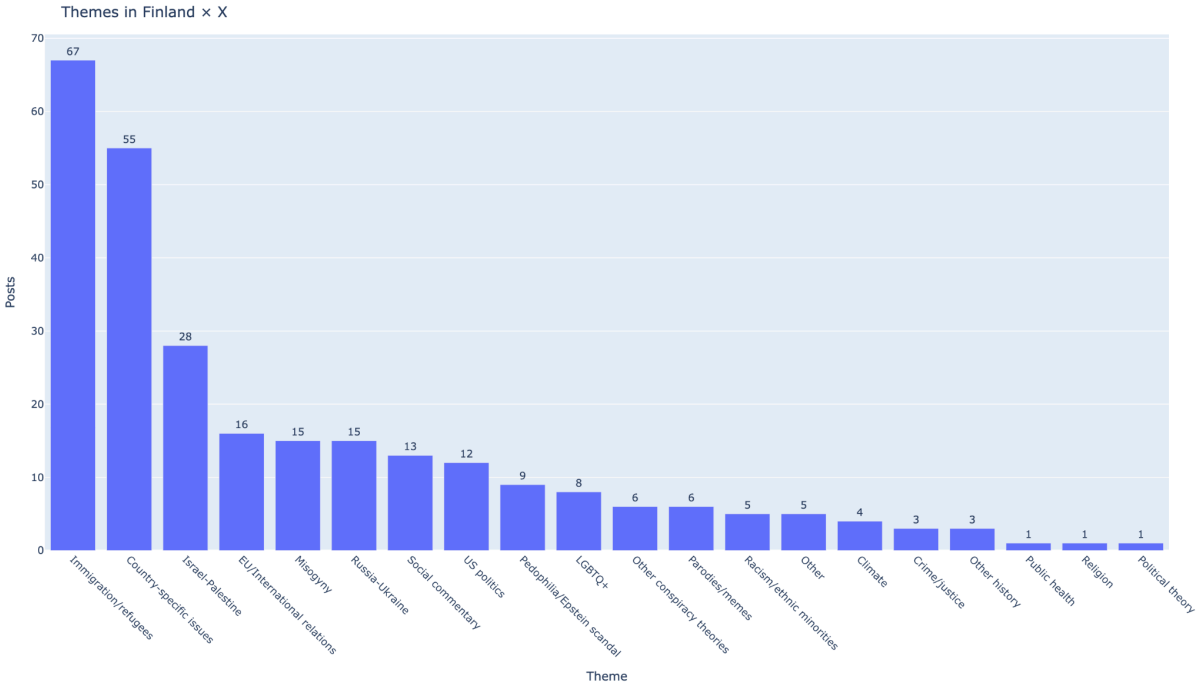


X

The journeys on X were characterised by a constant, high level of political content, especially posts sharing right-wing views. This did not change substantially across various sessions. Notably, problematic content, such as hate speech, misinformation and conspiracy theories were especially prominent in the first browsing sessions, shortly after the avatars signing up to the platform.

The most prominent themes were immigration and refugees and country-specific issues (some of which also related to immigration). Immigration was primarily discussed in a negative light, from a right-wing perspective. In contrast, the third most frequently discussed theme, the Israel-Palestine conflict was typically discussed from a left-wing perspective, heavily criticising Israel's actions. See Figure 18 for further details on the prominence of various themes across the two journeys on X.

Figure 18: Thematic analysis of all political posts (Finland, X)



Right-wing avatar

This user journey (see Figures 19 and 20) had a consistently high level of political content (note that the second Low Engagement session was extended to accommodate for an issue with language settings in the first session). The journey contained six conspiracy theories, four instances of hate speech and a few additional posts containing misinformation or hostile speech. These posts appeared at the beginning and at the end of the journey, with the middle sessions containing no such content. The feed displayed both right-wing and left-wing content, with right-wing being more prominent in some sessions. The journey covered a broad range of themes, but there were two notable patterns. First, right-wing posts often focused on immigration to Finland and sometimes contained hateful or hostile phrasings or shared related conspiracy theories. Second, left-wing posts focused most frequently on the Israel-Palestine conflict, strongly criticising Israel, and other international conflicts from history to the present day.

Figure 19: User journey illustration (Finland, X, Right-wing avatar)

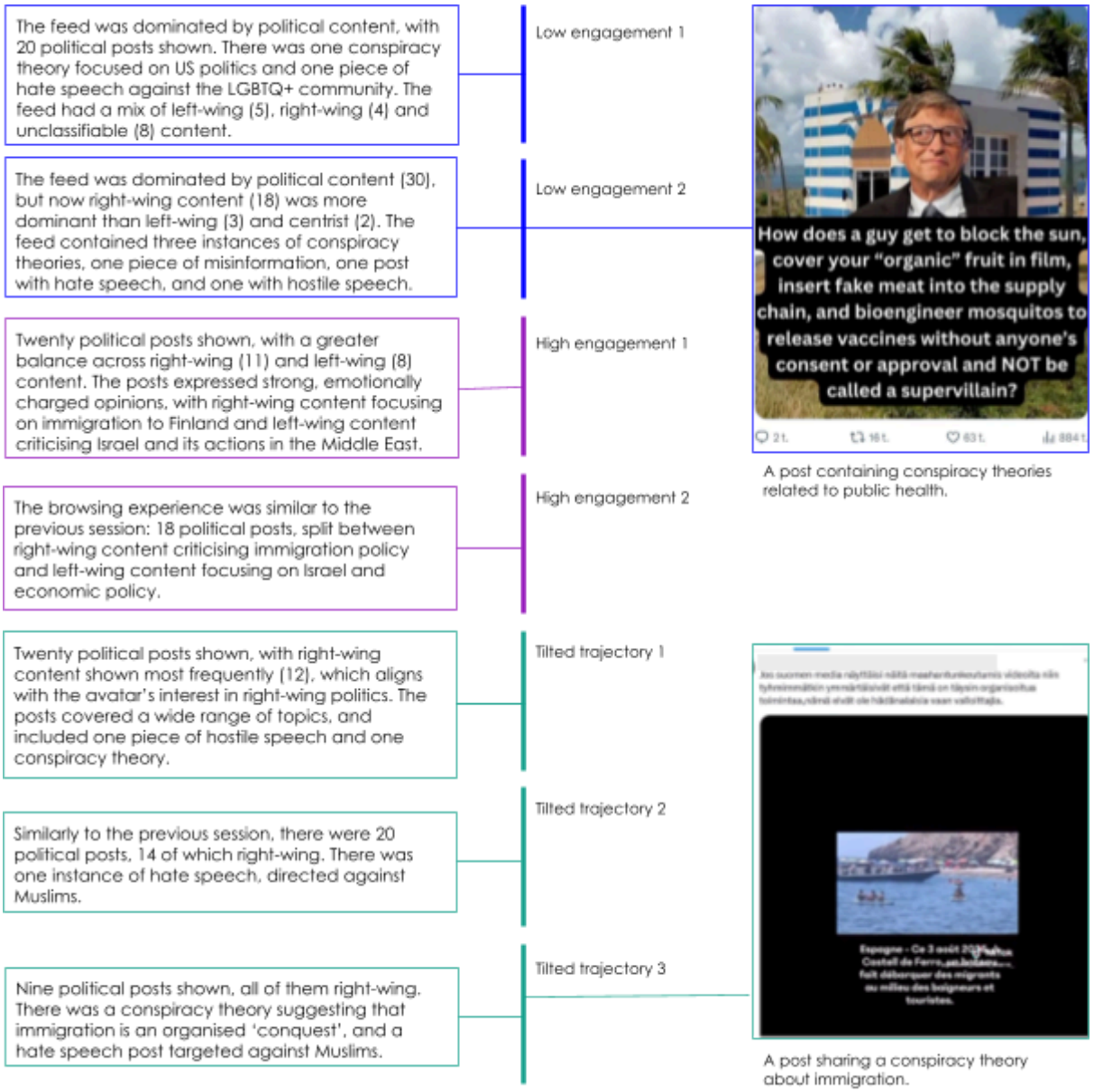
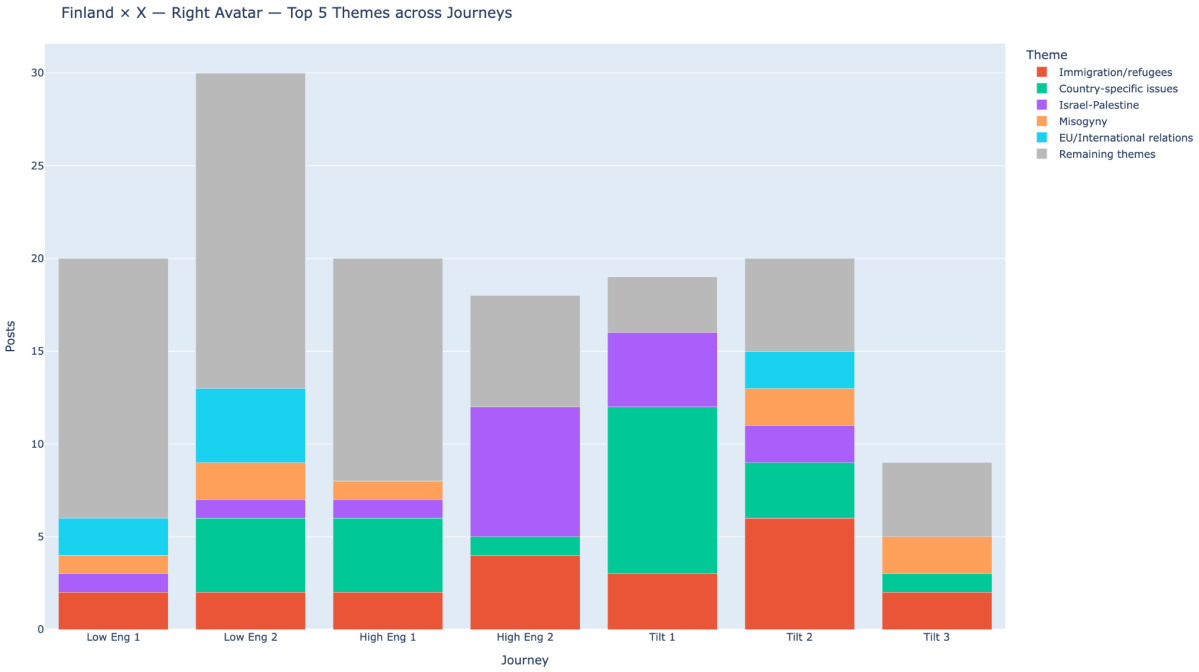


Figure 20: Thematic analysis of political posts encountered in user journey (Finland, X, Right-wing avatar)



Left-wing avatar

This user journey (see Figures 21 and 22) had a consistently high level of political content (note that the second Low Engagement session was extended to accommodate for an issue with language settings in the first session). The initial sessions each contained several pieces of misinformation, conspiracy theories, and hostile speech, mostly directed against Muslims or immigrants. The volume of such problematic content gradually decreased as the journey progressed. Overall, the avatar saw one hate speech post, 14 posts with hostile speech, eight pieces of misinformation and two conspiracy theories. Throughout the entire journey, right-wing content was more frequent than left-wing or centrist content, despite the avatar expressing interest in left-wing politics for the last three browsing sessions. The posts discussed immigration and local issues, as well as international themes, such as the Israel-Palestine conflict and EU policies.

Figure 21: User journey illustration (Finland, X, Left-wing avatar)

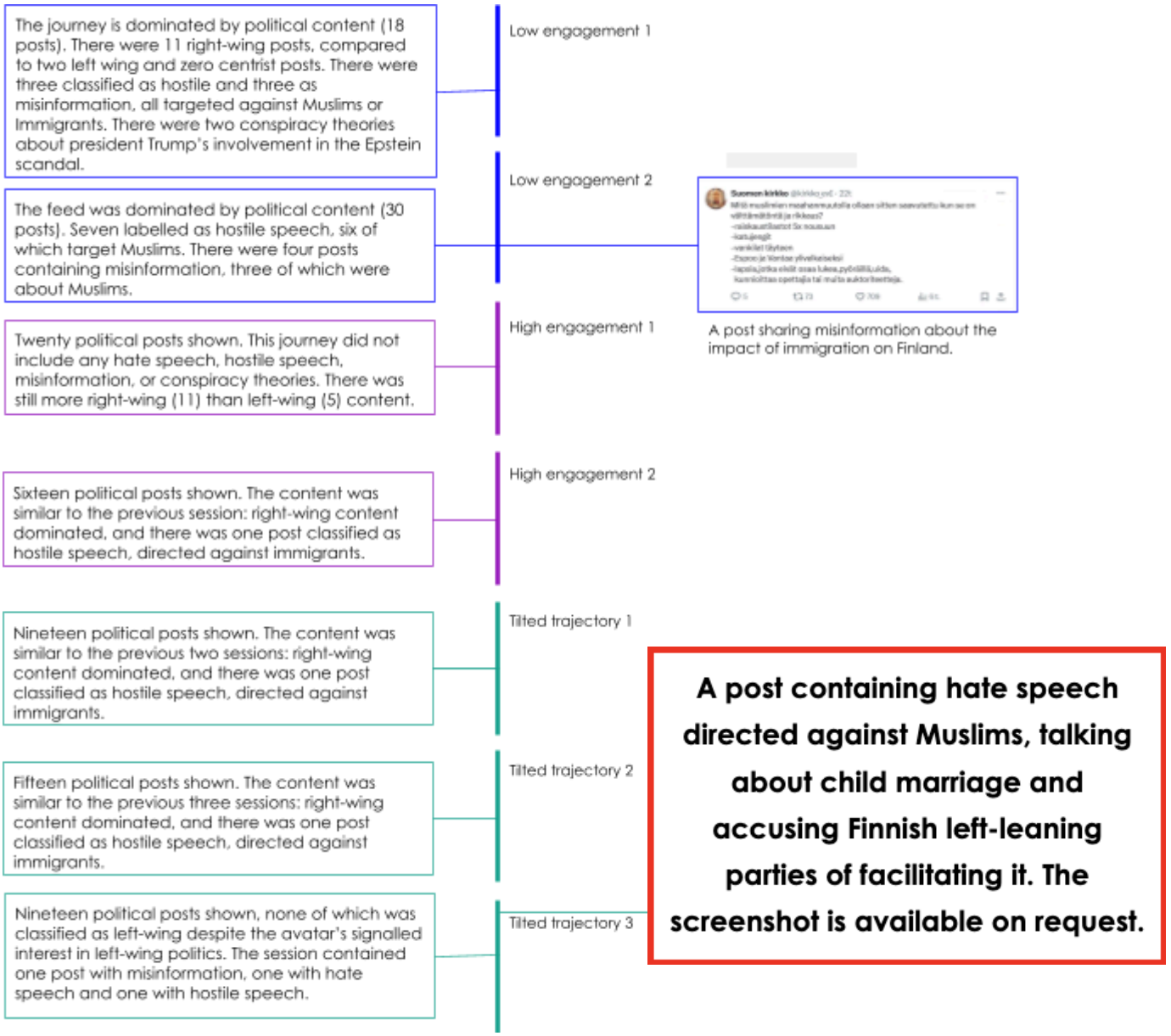
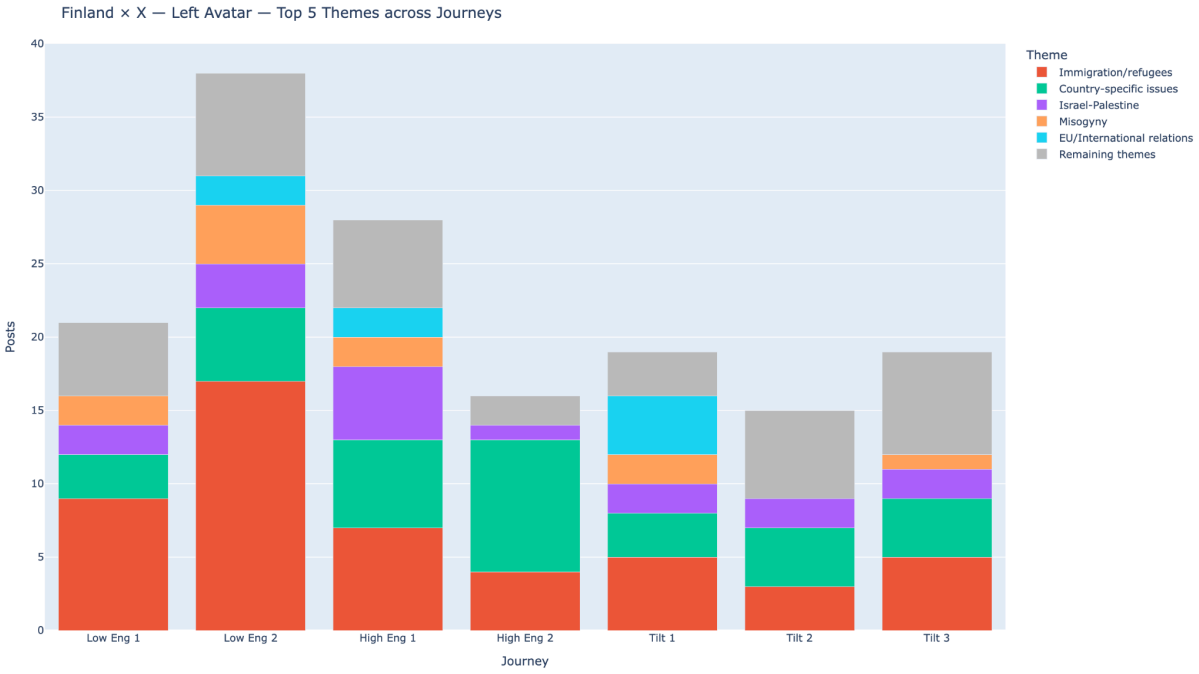


Figure 22: Thematic analysis of political posts encountered in user journey (Finland, X, Left-wing avatar)

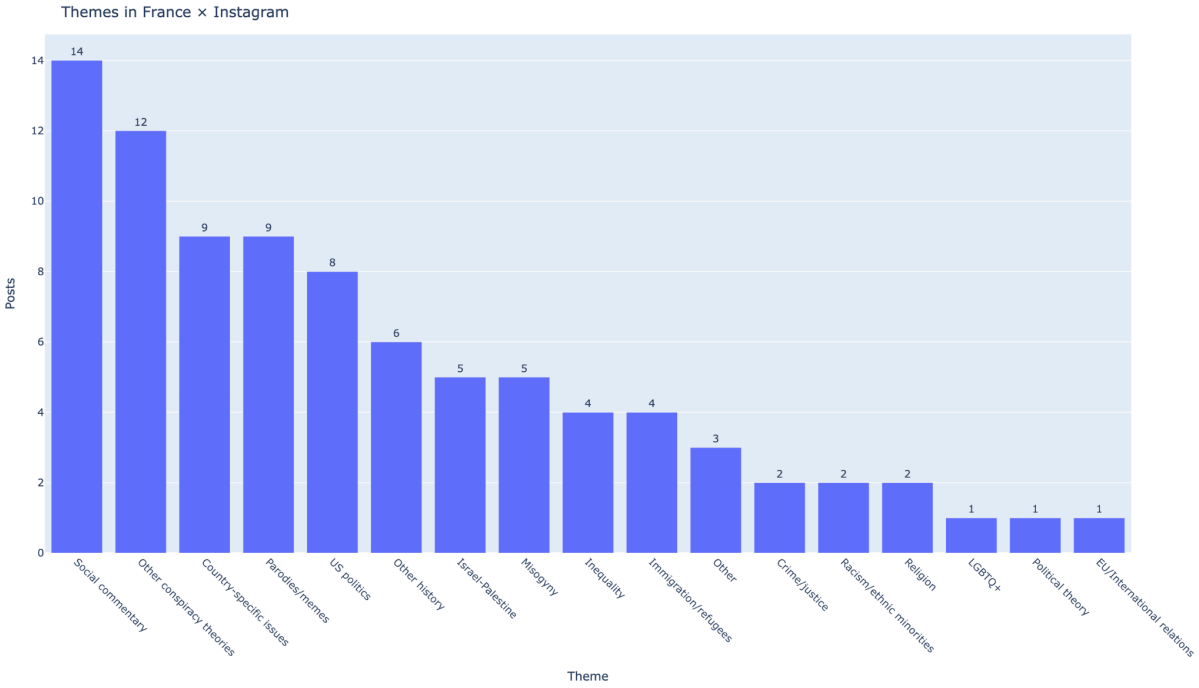


France

Instagram

The political content encountered by the avatars on Instagram included a mix of right- and left-wing posts, although right-wing material was most frequently displayed. Few centrist posts were observed, and levels of misinformation, malinformation, and conspiracy theories were low, with no instances of hate or hostile speech. Political content generally peaked when the avatars expressed interest in politics, though tilts in preference had a somewhat counterintuitive effect: left-tilted activity was associated with an increase in right-wing posts, and right-tilted activity corresponded with a rise in left-wing posts. Across both journeys, recurring themes included social commentary — that is, reflections on cultural issues —, conspiracy theories not captured by other themes, country-specific issues, and parodies or memes (see Figure 23 for the thematic breakdown of posts).

Figure 23: Thematic analysis of all political posts (France, Instagram)



Right-wing avatar

On this user journey (see Figures 24 and 25), the avatar encountered some misinformation, malinformation, and conspiracy theories, but no hostile content or hate speech. The volume of political posts peaked when the avatar expressed interest in political content. Right-wing content was minimal, while left-wing content remained low initially and increased once right-wing preferences were expressed. No centrist content was observed. Prominent themes in this user journey included conspiracy theories, parodies and memes, against a variety of themes.

Figure 24: User journey illustration (France, Instagram, Right-wing avatar)

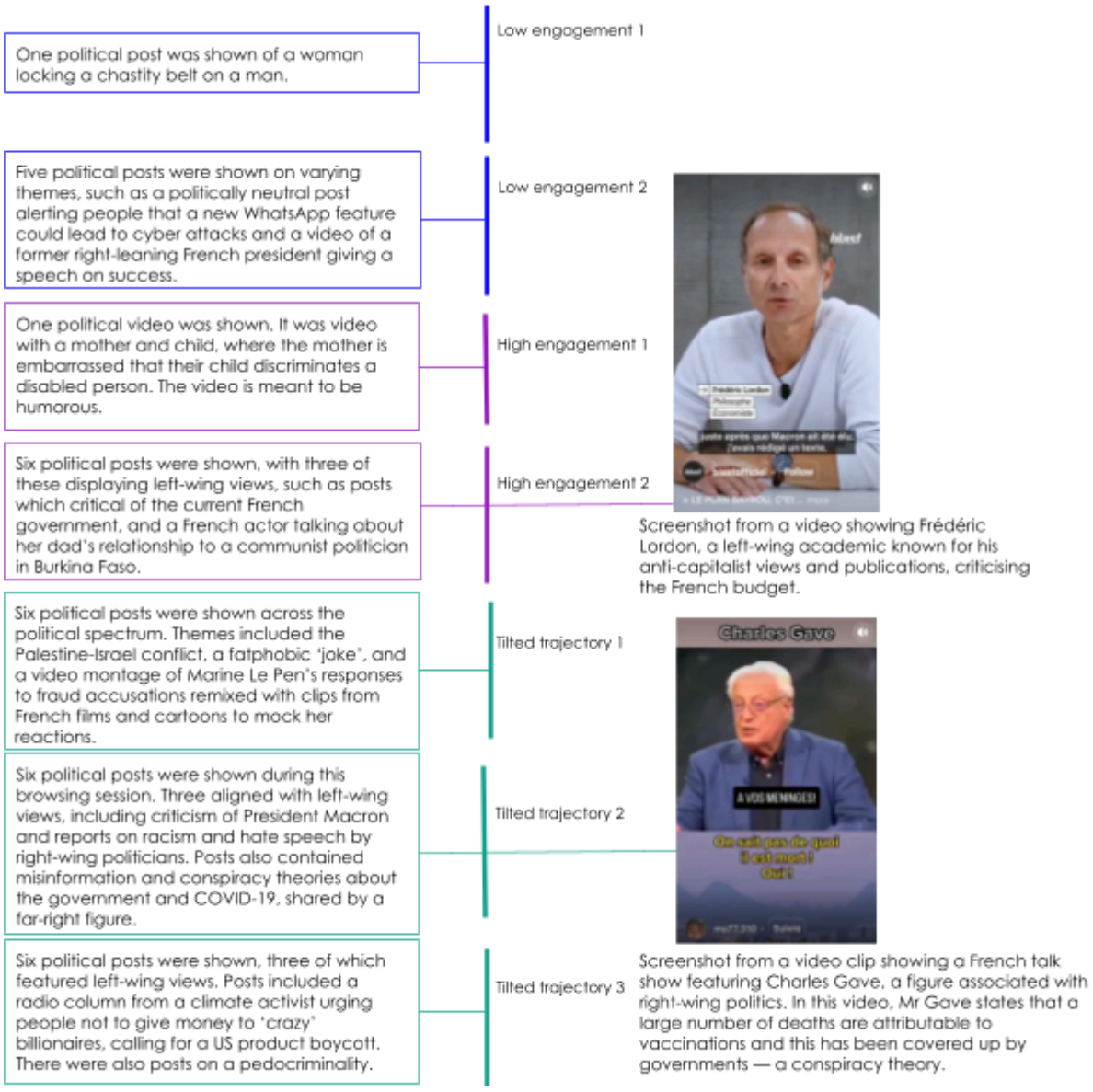
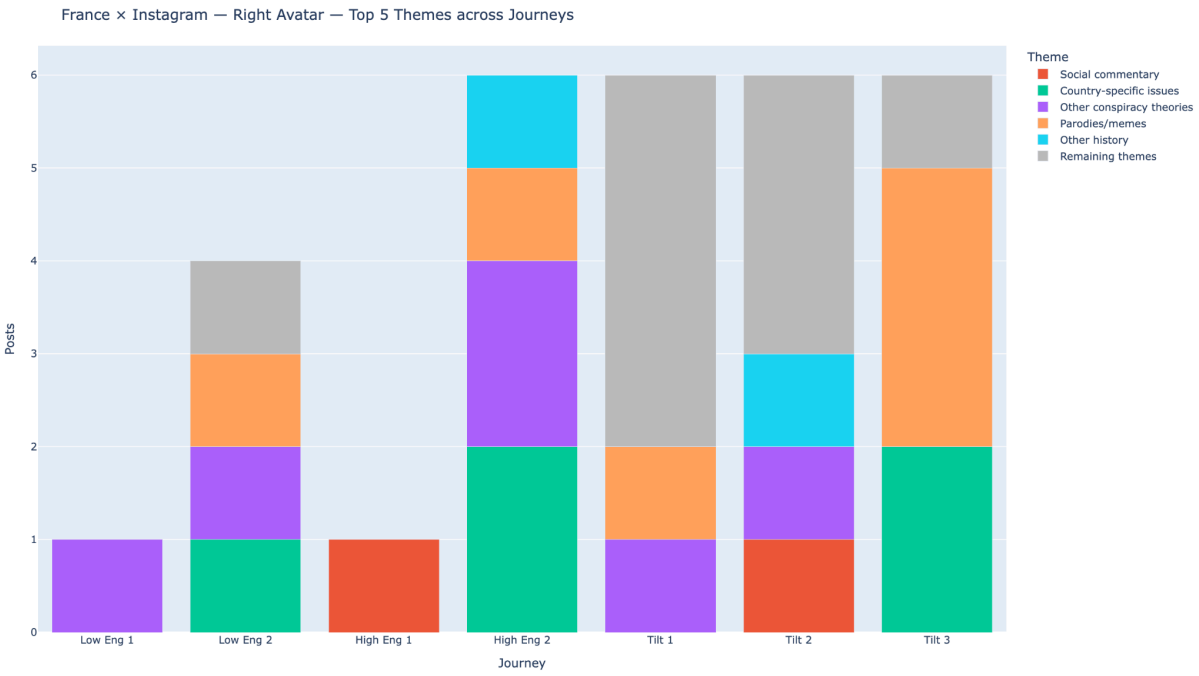


Figure 25: Thematic analysis of political posts encountered in user journey (France, Instagram, Right-wing avatar)



Left-wing avatar

On this journey (see Figures 26 and 27), the avatar encountered some misinformation, malinformation, but no hostile content or hate speech. Political posts were moderately represented, peaking when the avatar expressed interest in political content. Right-wing content was most prominent, even after the avatar indicated a preference for left-wing political content. Left-wing content remained consistently low and centrist content was rare. Prominent themes include country-specific issues, posts centered on social commentary, and other conspiracy theories.

Figure 26: User journey illustration (France, Instagram, Left-wing avatar)

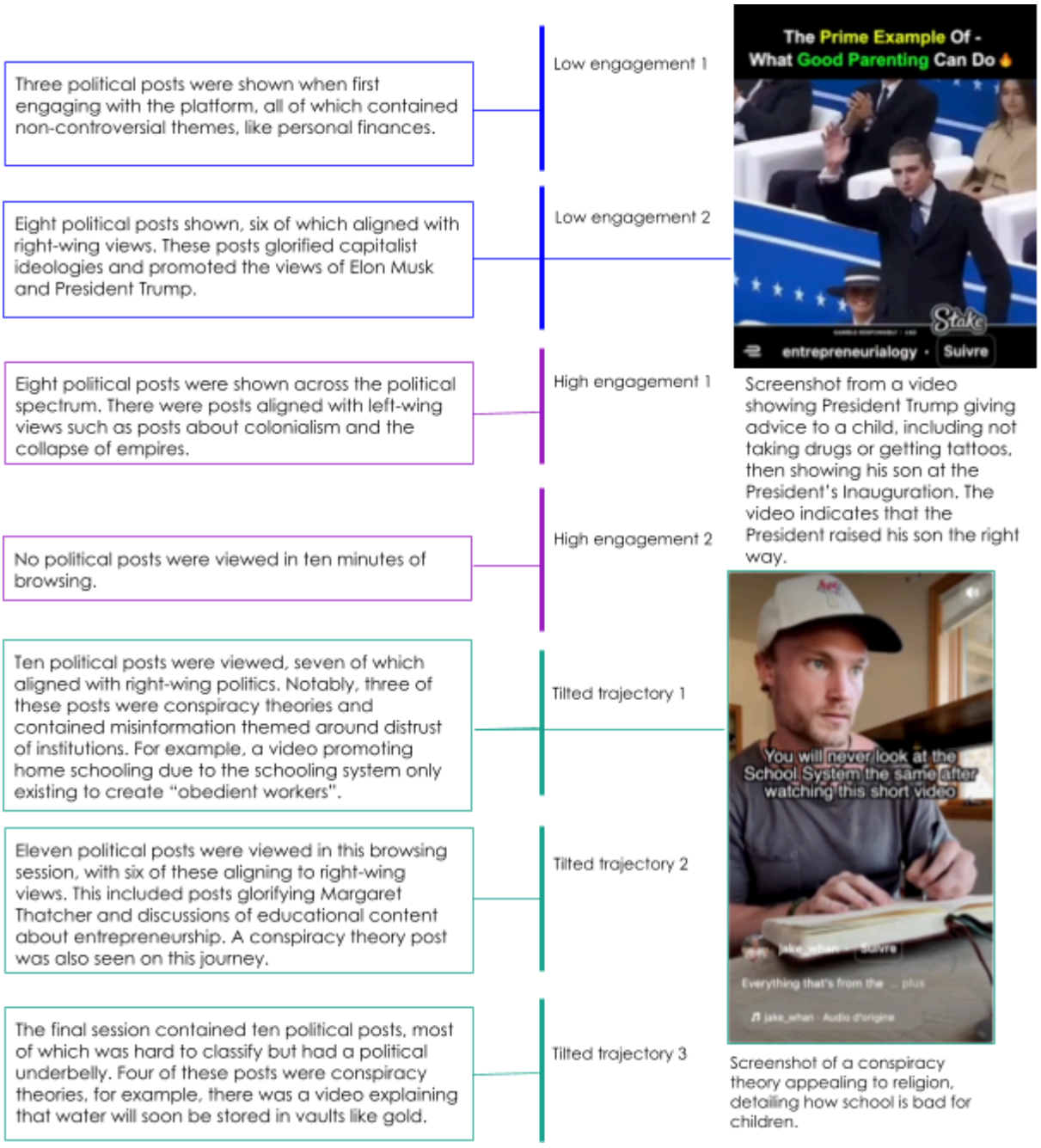
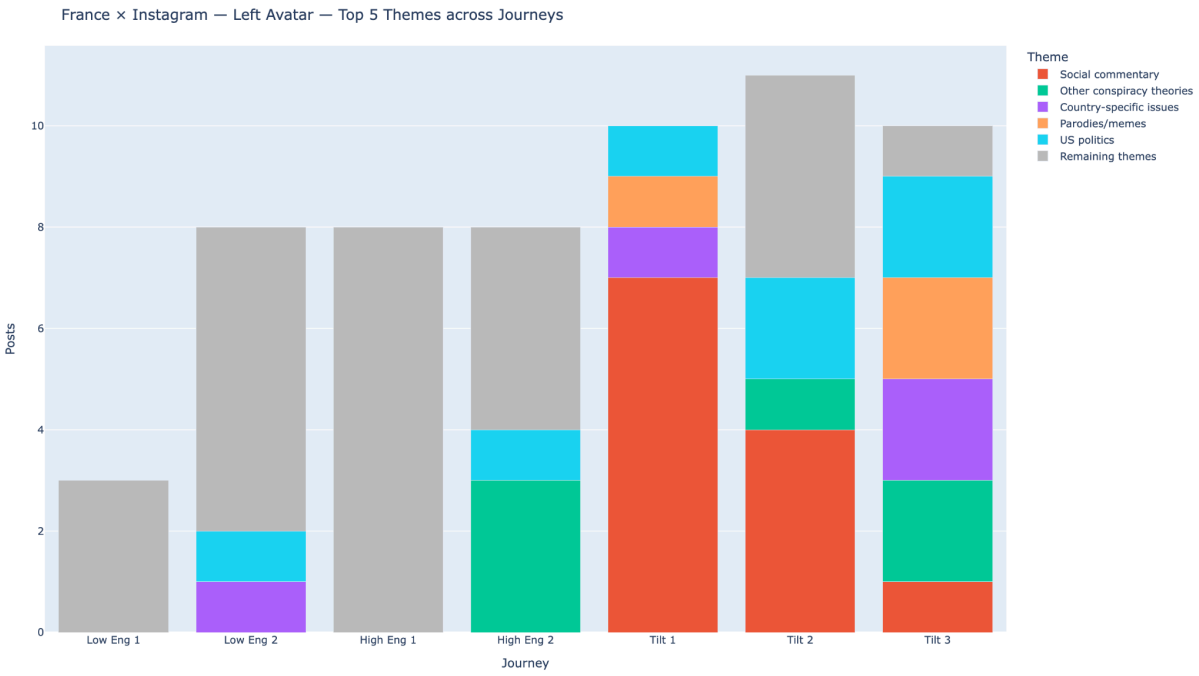


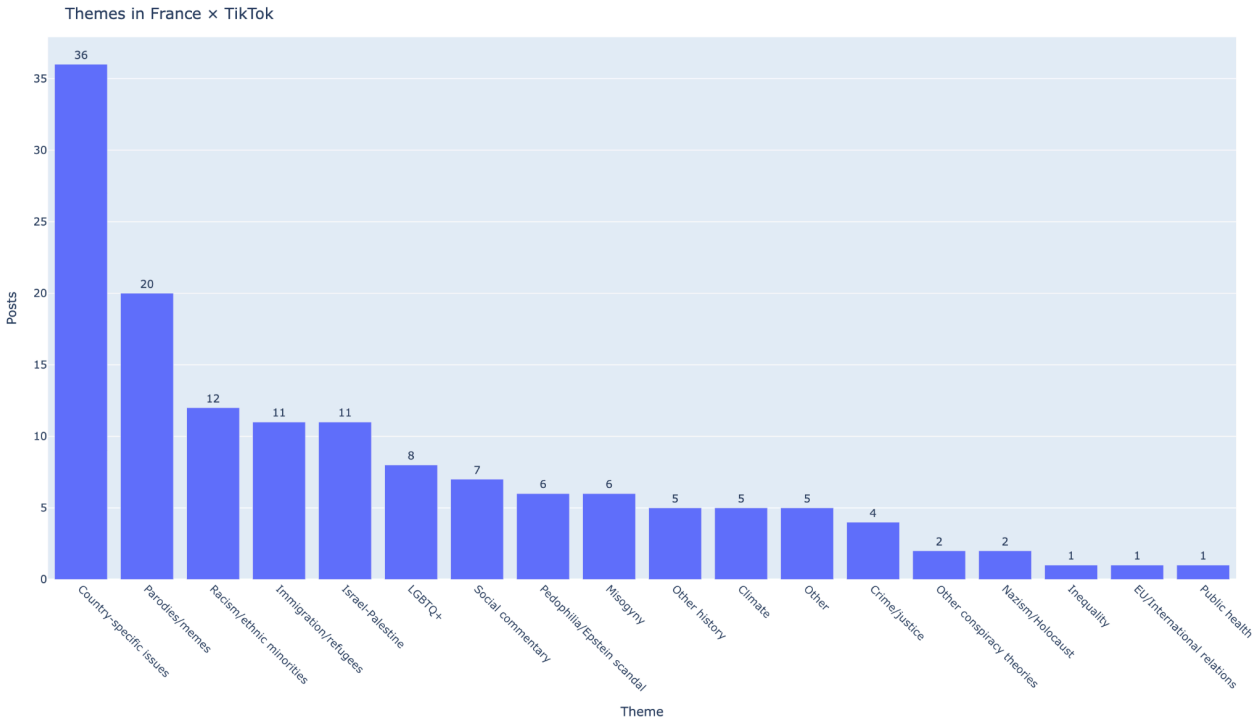
Figure 27: Thematic analysis of political posts encountered in user journey (France, Instagram, Left-wing avatar)



TikTok

The political content encountered by the avatars on TikTok was dominated by right-wing material, with left- and centrist-leaning posts appearing less frequently. Thematic patterns were broadly consistent across both journeys, featuring posts on country specific issues, immigration, ethnic minorities, and parodies and memes (see Figure 28 for the thematic breakdown of posts). A small number of posts contained conspiracy-type content, misinformation, or hate and hostile speech, most of which appeared early in the browsing sessions. Overall, levels of political and problematic content remained relatively constant, and the avatars' differing political orientations had little effect on the content displayed.

Figure 28: Thematic analysis of all political posts (France, TikTok)



Right-wing avatar

On this user journey (see Figures 29 and 30), the avatar encountered a moderate amount of political content, highest at initial engagement before settling to a consistent lower level. Right-wing content dominated throughout, while left-wing and centrist posts were limited. One post with conspiracy theories was observed, but no broader misinformation was identified. Some hostile content and one instance of hate speech appeared early in the journey, decreasing as engagement continued. During this user journey there were many posts on country-specific issues which included policy discussions and speeches by right-wing EU representatives. Other popular themes included parodies/memes and the Israel-Palestine conflict.

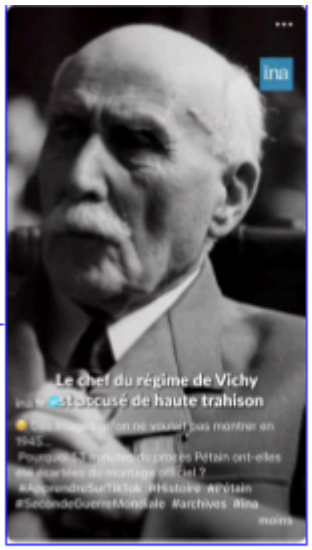
Figure 29: User journey illustration (France, TikTok, Right-wing avatar)

Eighteen political posts were shown when first engaging with the platform, which ranged across the political spectrum. This part of the journey featured posts aligning to right-wing narratives with racist and islamophobic content, often intended to be humorous.

Low engagement 1

Twelve political posts shown with varying themes, for example the Israel-Palestine conflict, the Nazi occupation of France and body positivity.

Low engagement 2



Screenshot of a documentary-style video of a prior French president who was accused of collaboration with the Nazis in 1945.

Seven political posts were shown, one of which included anti-non-binary hostile speech, claiming that non-binary people need psychological help.

High engagement 1

There were six political posts with a mix of political views and themes. There was a video of an interview with a right-leaning politician asking viewers whether they agree with the statements made, posts on immigration and a post of an anti-capitalist speech given in Parliament.

High engagement 2

Six political posts were shown, some of which were of a harsh subject matter. Examples included politically neutral posts detailing the reality of being a drug mule, a post surrounding sexual abuse and a post directed against non-binary people.

Tilted trajectory 1

Screenshot of an aggressive post where a man belittles a complaint made by a non-binary person about there being no non-binary toilets in the Louvre.

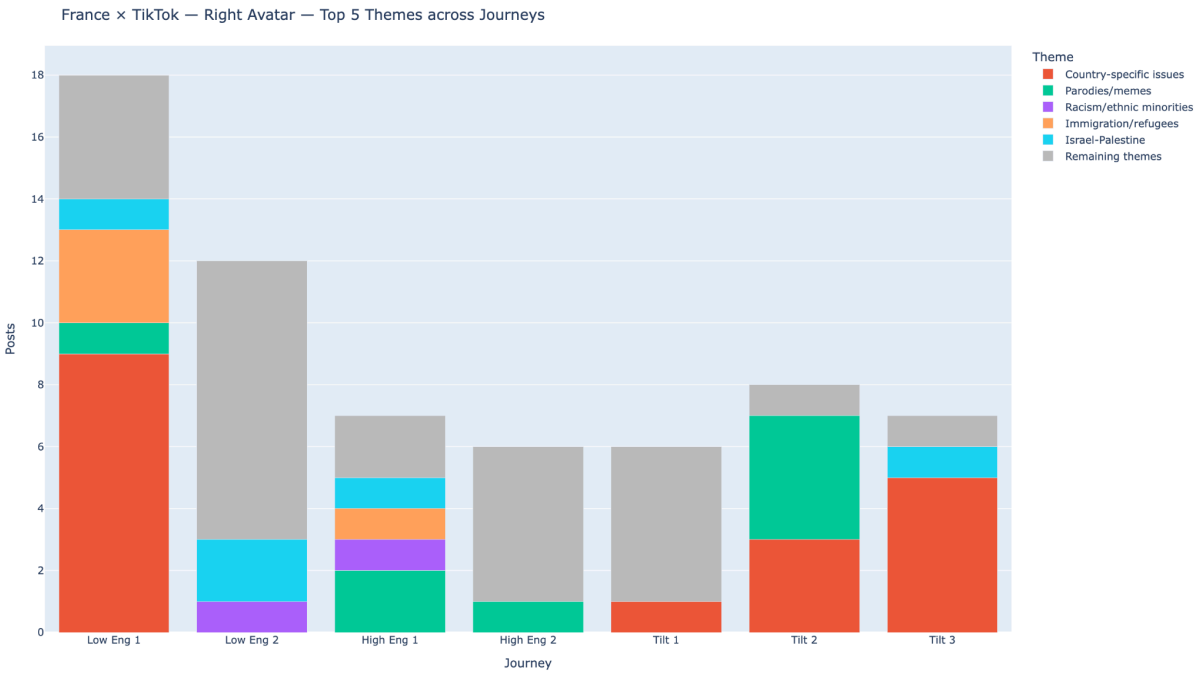
Eight political posts were viewed in this browsing session, with six of these focusing on French local and party politics. For example, a humorous post featured clips of French politicians from across the political spectrum.

Tilted trajectory 2

The final session contained seven political posts, five of which aligned with right-wing political views. For example, one post advocated that France should establish national solidarity and that family benefits must be reserved for French families.

Tilted trajectory 3

Figure 30: Thematic analysis of political posts encountered in user journey (France, TikTok, Right-wing avatar)

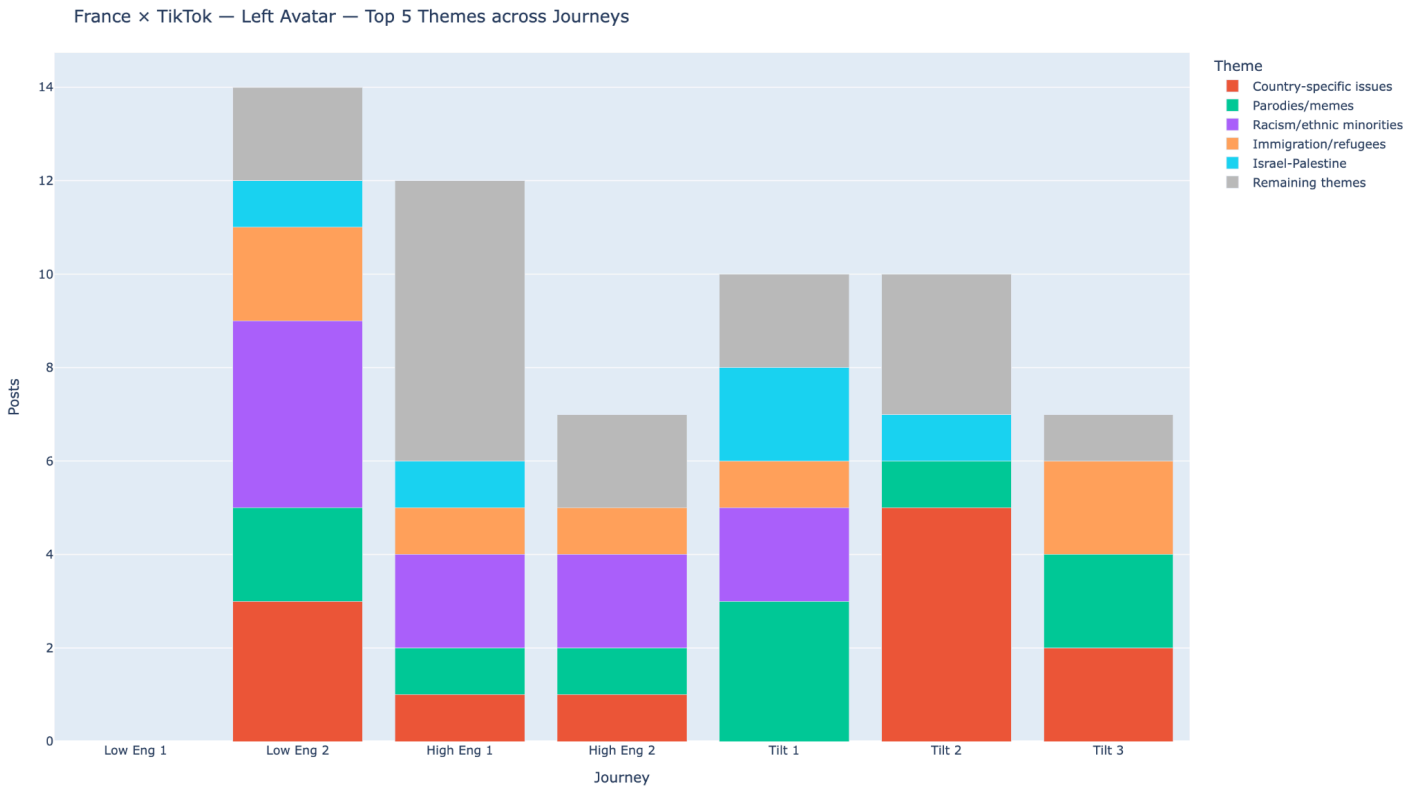


Left-wing avatar

On this user journey (see Figures 31 and 32), the avatar encountered mostly right-wing political content, particularly at the start, with lower levels of left-wing and centrist posts throughout. Only one conspiracy-type post was observed, and no broader misinformation was identified. Some hostile content and a single instance of hate speech appeared early in the journey, with hostile content increasing slightly as engagement grew. Prominent themes in this user journey included racist posts or those focused on ethnic minorities, immigration, parodies/memes and country-specific issues.

Figure 31: User journey illustration (France, TikTok, Left-wing avatar)

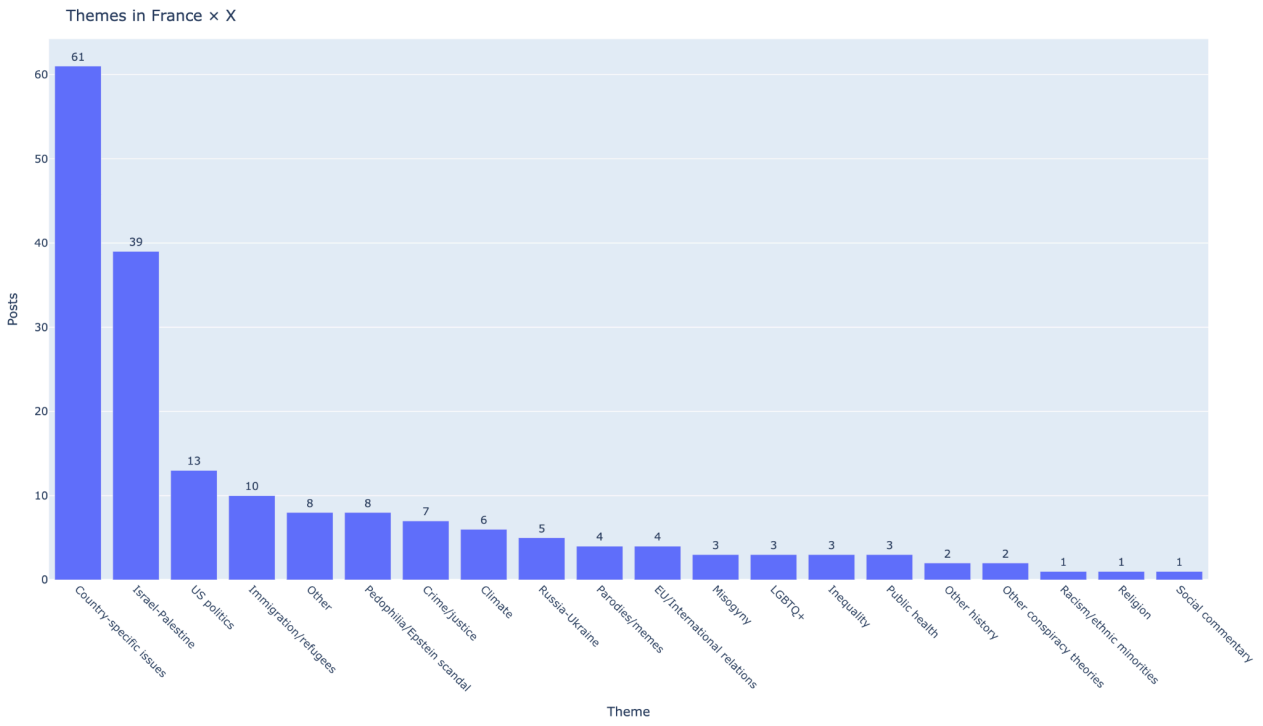
Figure 32: Thematic analysis of political posts encountered in user journey (France, TikTok, Left-wing avatar)



X

The political content encountered by the avatars on X spanned a wide range of themes and orientations. Contrary to most other journeys, left-wing material appeared most frequently. Posts commonly addressed country-specific issues and international political issues, particularly US politics and the Israel–Palestine conflict (see Figure 33 for the thematic breakdown of posts). Misinformation, conspiracy theories, and hate or hostile speech appeared only occasionally. Broadly, the overall volume of political content declined during the user journey.

Figure 33: Thematic analysis of all political posts (France, X)



Right-wing avatar

On this user journey (see Figures 34 and 35), the avatar initially encountered a high level of conspiracy theories, which declined substantially over time. Only a few posts contained mis- or malinformation. One instance of hate speech was observed, with no other hostile content. At first engagement, right-wing content was moderately represented, decreasing as the user engaged more, where left-wing content remained dominant throughout. This change was contrary to our broader findings that user signals did not change feeds. Across the journey, prominent themes included the Israel-Palestine conflict and country-specific issues, alongside a wide variety of other topics.

Figure 34: User journey illustration (France, X, Right-wing avatar)

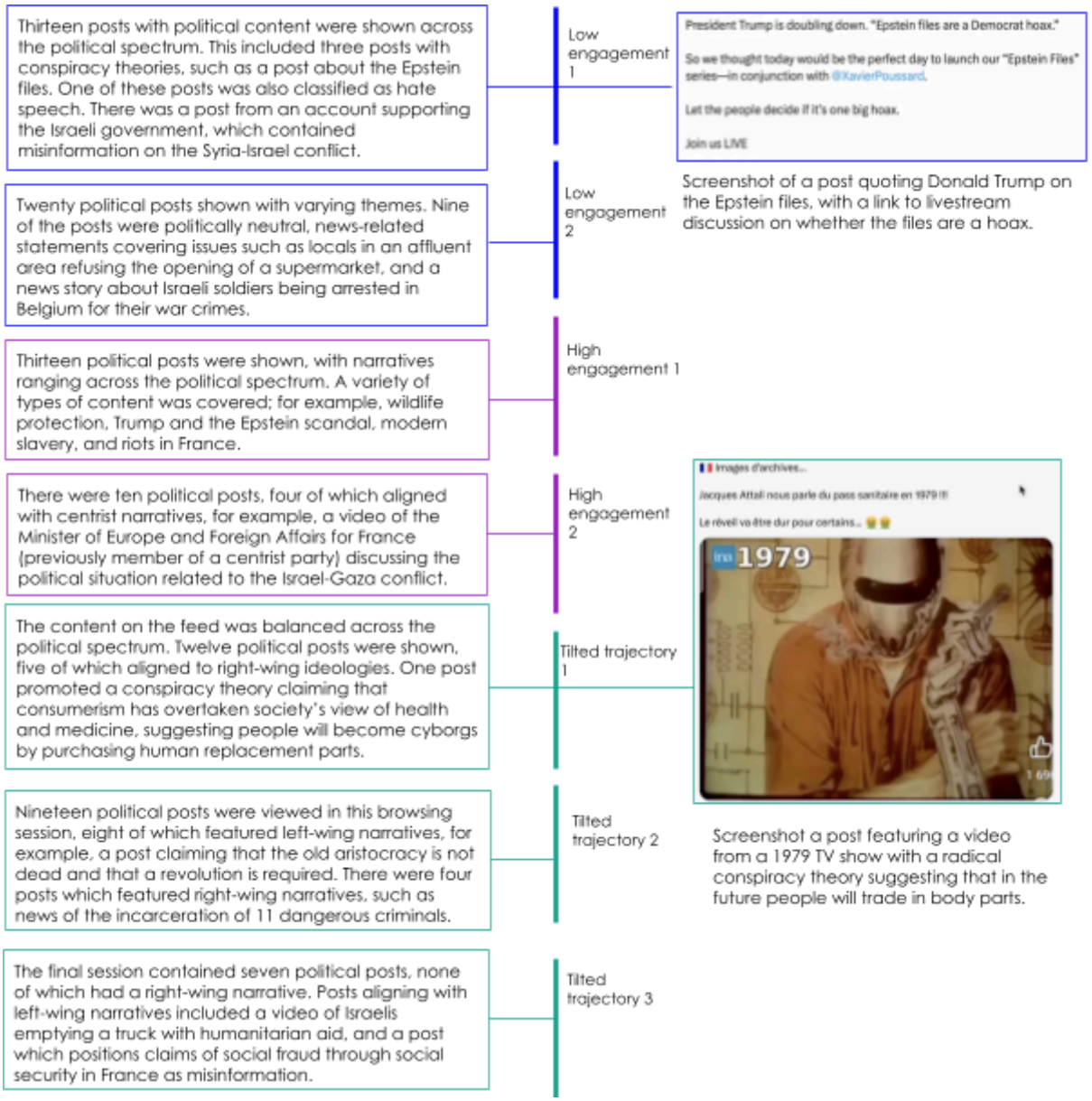
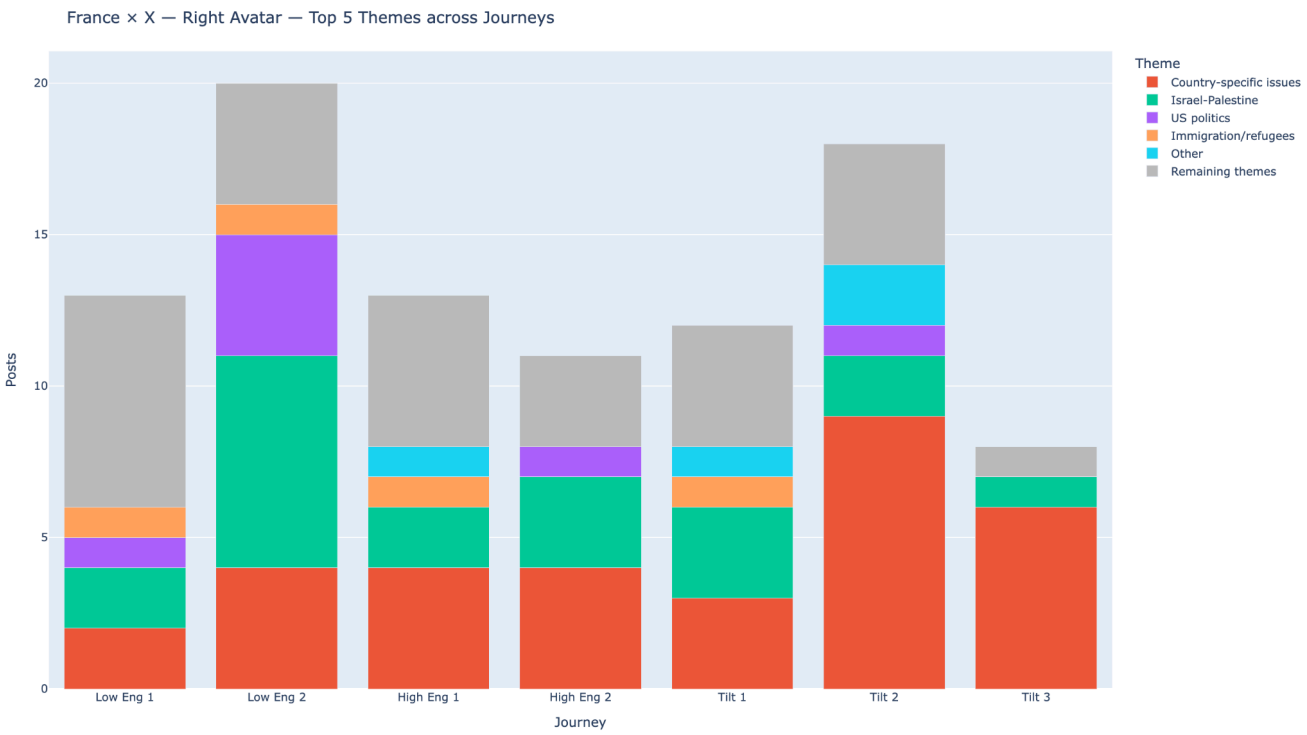


Figure 35: Thematic analysis of political posts encountered in user journey (France, X, Right-wing avatar)



Left-wing avatar

On this user journey (see Figure 36), the avatar saw many political posts, but only a few of these contained mis- or malinformation or conspiracy theories. None of the posts contained hate speech or hostile speech. Initially, there was a mix of left-wing, centrist, and right-wing content, covering a wide range of themes. Over time, the total amount of political content decreased, despite increased interest from the avatar. Left-wing content became more dominant towards the final sessions, in line with the avatar's stated interest in left-wing politics. Across all stages of the user journey the avatar saw posts on the Israel-Palestine conflict, often reflecting a pro-Palestine perspective. Other popular post themes include posts relating to immigration or refugees, and country-specific issues.

Figure 36: User journey illustration (France, X, Left-wing avatar

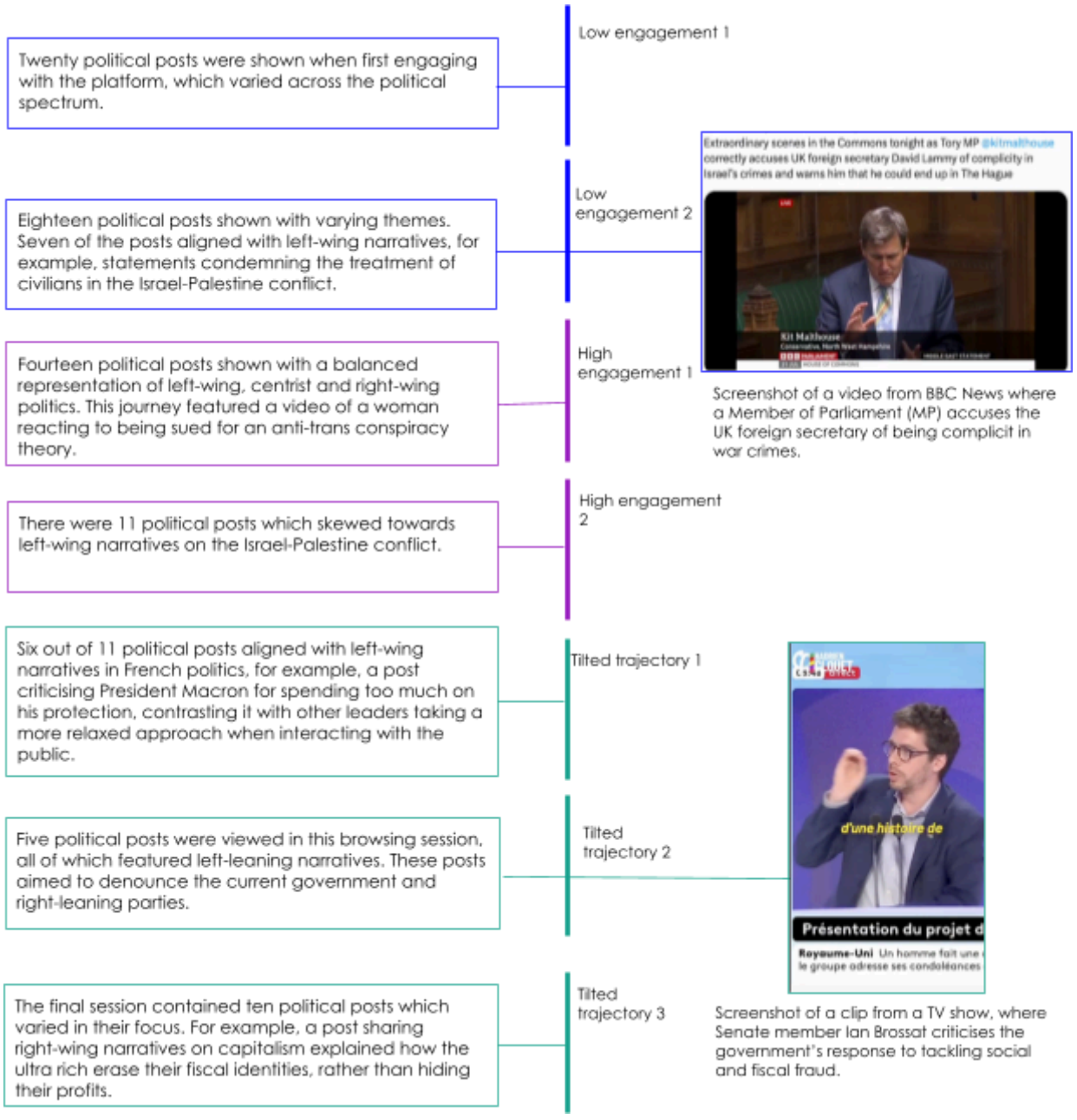
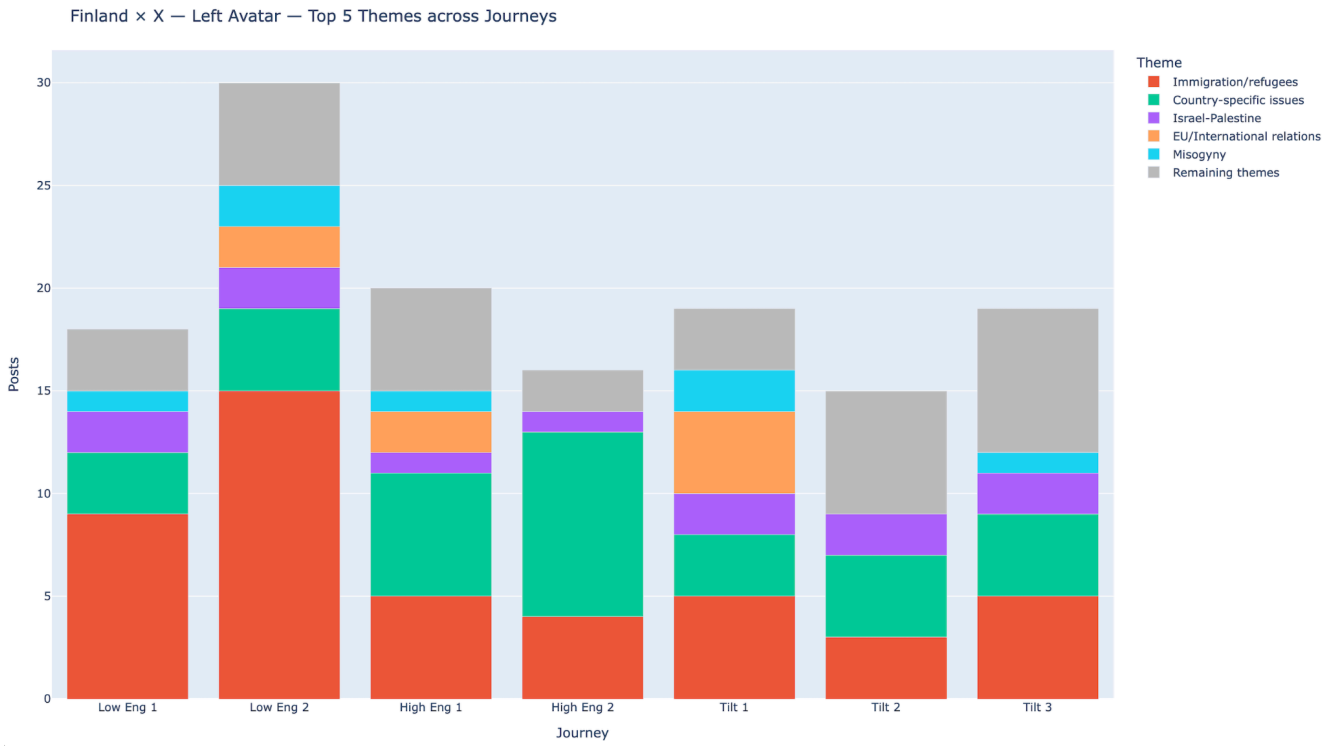


Figure 37: Thematic analysis of political posts encountered in user journey (France, X, Left-wing avatar

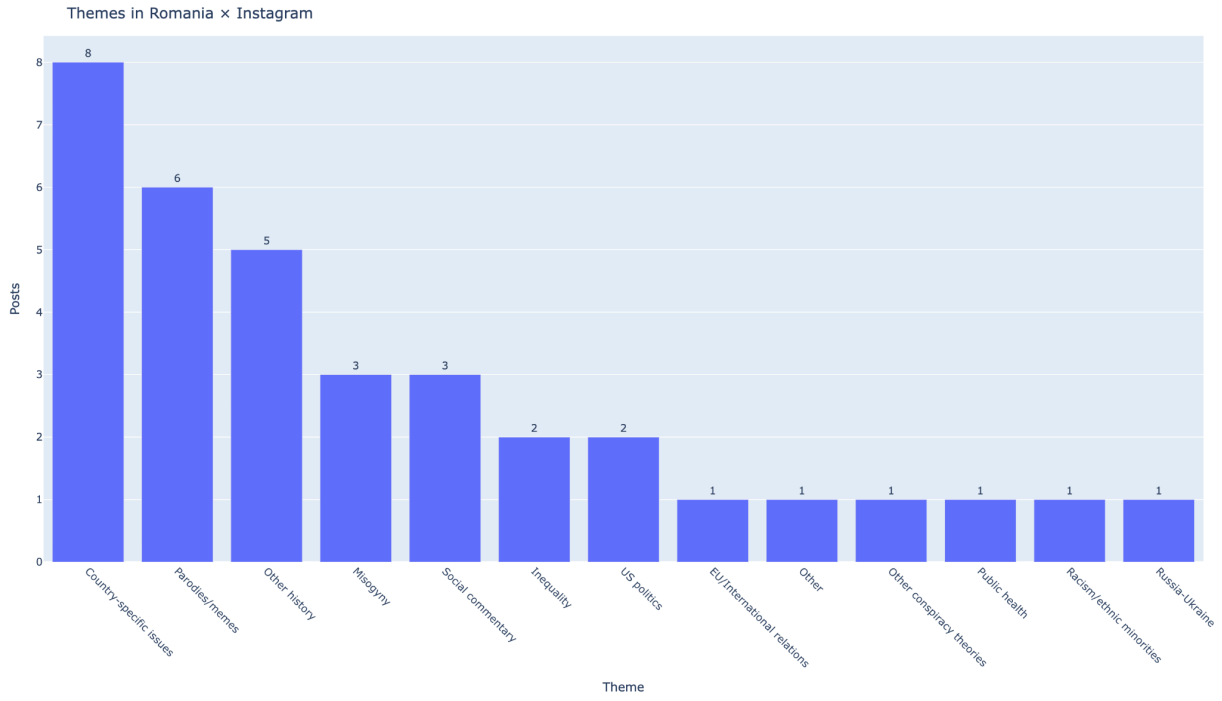


Romania

Instagram

On Instagram in Romania, political content was generally low in volume across both avatars, with right- and left-wing posts appearing infrequently and centrist posts limited. Levels of conspiracy theories and misinformation were low, and no hate speech was observed; a single instance of hostile content appeared during the left-avatar journey. Political posts were most prominent during the initial browsing sessions, particularly for the left-wing avatar, which initially encountered a higher proportion of right-wing posts. Recurring themes included country-specific issues, other history, and parodies or memes, including satirical posts about past leaders (see Figure 38 for the thematic breakdown of posts).

Figure 38: Thematic analysis of all political posts (Romania, Instagram)



Right-wing avatar

On this user journey (see Figures 39 and 40), the avatar encountered some conspiracy theories and misinformation. However, no malinformation, hate speech, or hostile material was observed. This avatar encountered a low number of political posts, primarily appearing during initial engagement before political preferences were expressed. No centrist content was identified. Of the few political posts encountered, themes included social commentary and other history.

Figure 39: User journey illustration (Romania, Instagram, Right-wing avatar)

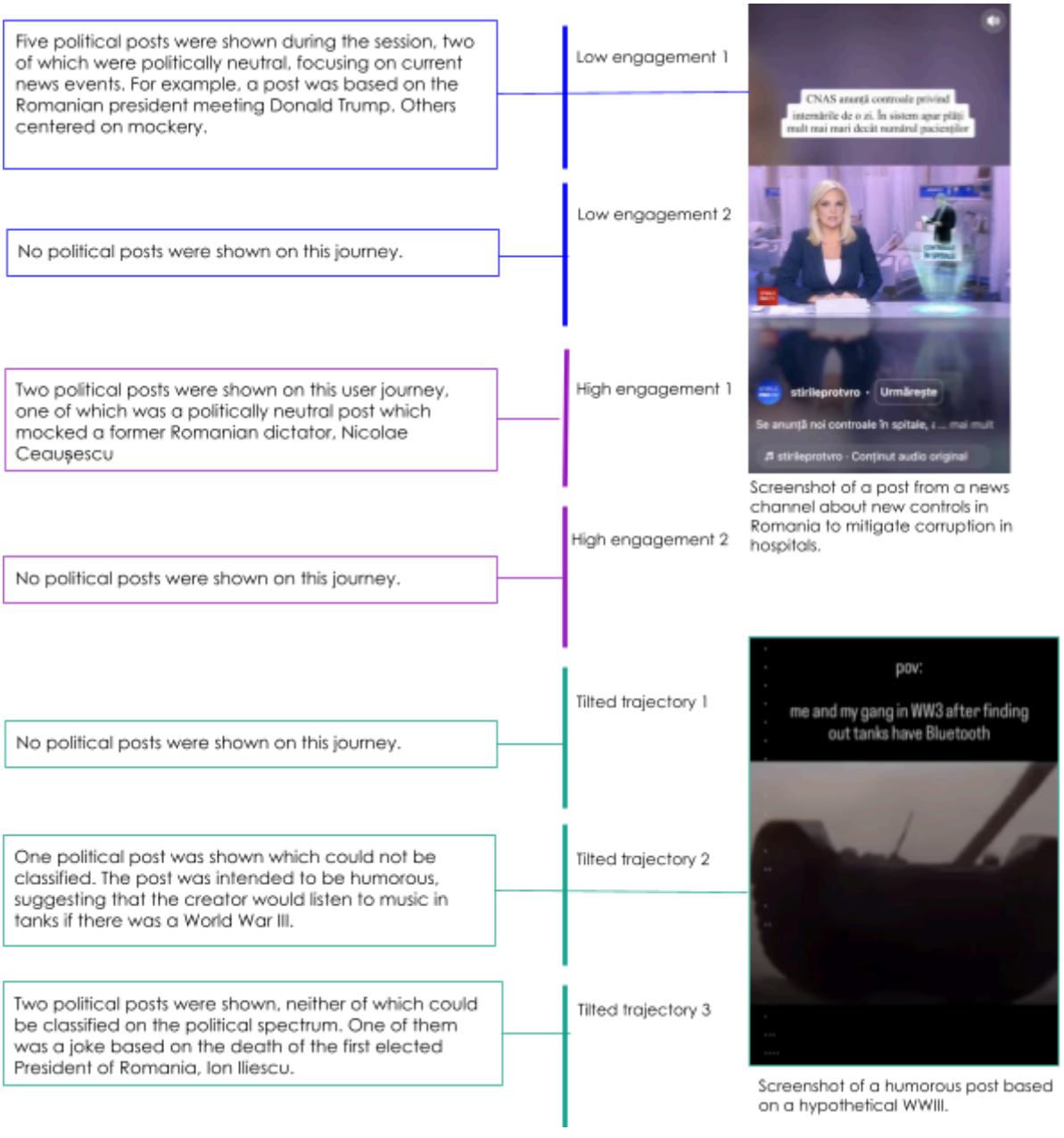
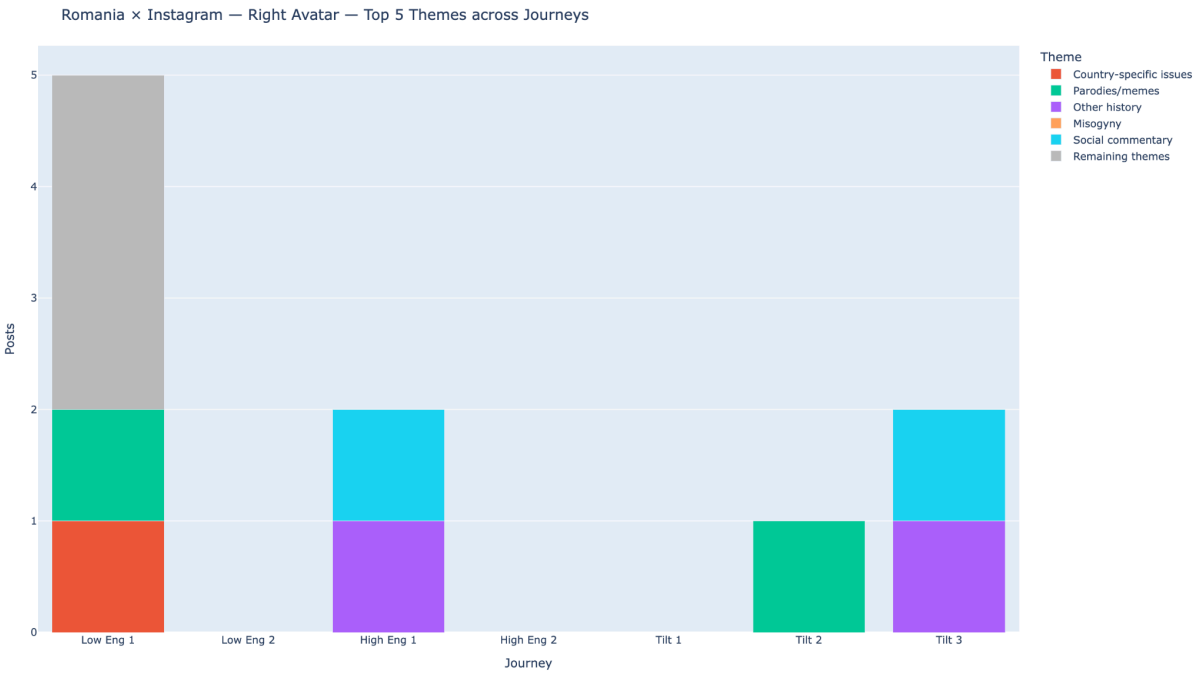


Figure 40: Thematic analysis of political posts encountered in user journey (Romania, Instagram, Right-wing avatar)



Left-wing avatar

On this user journey (see Figures 41 and 42), the avatar encountered some conspiracy theories and misinformation but no malinformation. Political posts were generally low in volume, with the exception of one session. Right-wing content was initially relatively high before decreasing when the avatar expressed left-wing preferences. This suggests that the algorithm reacted to user signals in this case. Left-wing content remained very low, and centrist posts were also limited. No hate speech was observed, though one piece of hostile content appeared. Themes included country-specific issues, historical posts, and parodies or memes, including satirical commentary on past leaders.

Figure 41: User journey illustration (Romania, Instagram, Left-wing avatar)

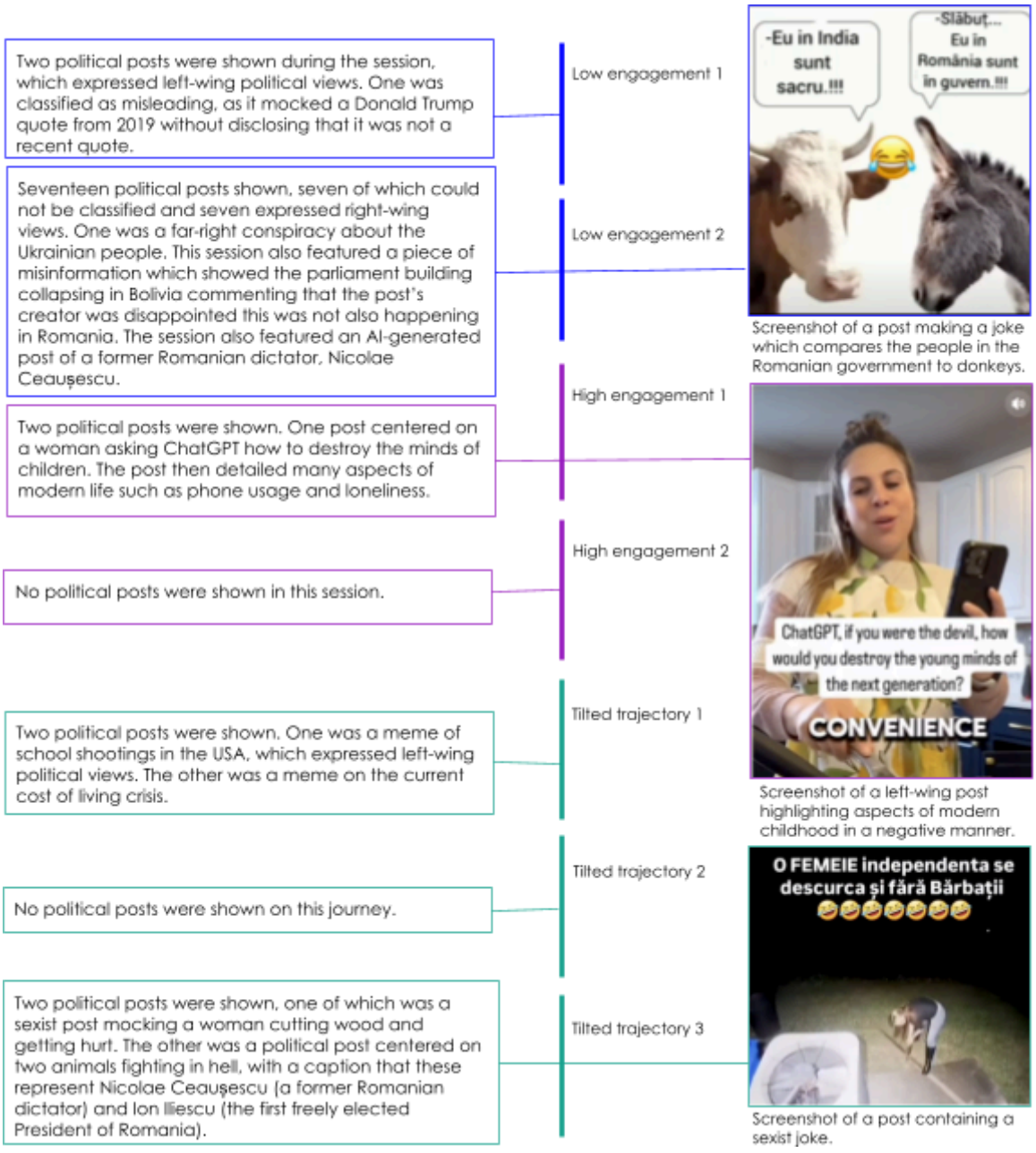
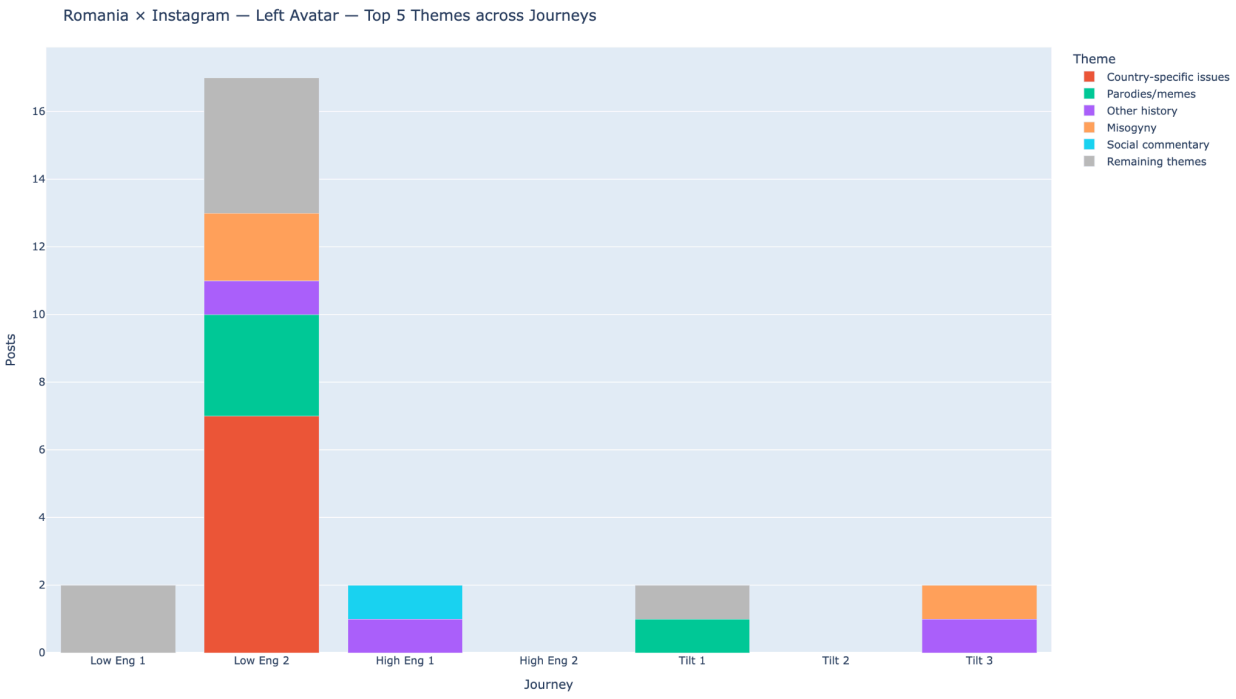


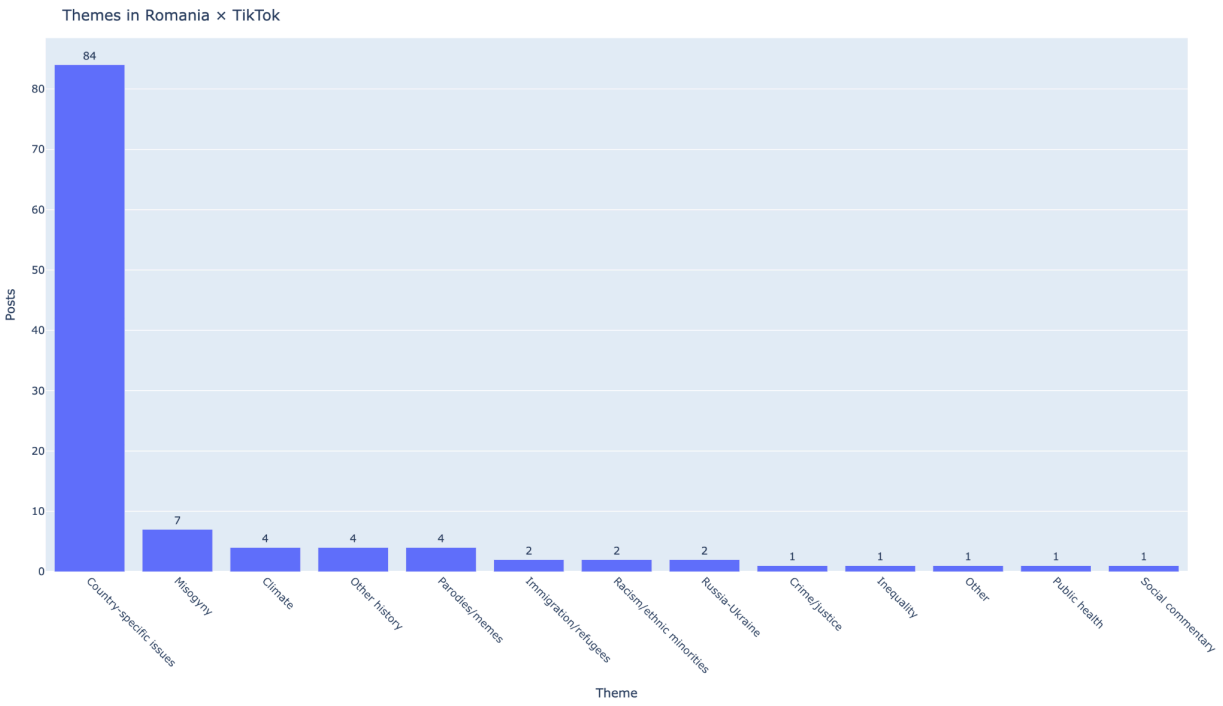
Figure 42: Thematic analysis of political posts encountered in user journey (Romania, Instagram, Left-wing avatar)



TikTok

On TikTok in Romania, political content was generally low in volume across both avatars, with right- and left-wing posts appearing infrequently and centrist content relatively more prominent, particularly for the right-wing avatar once preferences were expressed. Levels of misinformation, conspiracy theories, and malinformation were minimal, with only occasional hostile content and no hate speech observed. Political posts were largely dominated by country-specific issues, while other themes included misogyny, climate and other history (see Figure 43 for the thematic breakdown of posts).

Figure 43: Thematic analysis of all political posts (Romania, TikTok)



Right-wing avatar

On this user journey (Figures 44 and 45), the avatar did not encounter misinformation but saw conspiracy theories and some malinformation. Of the political posts encountered, right-wing content was low and absent once preferences towards right-wing political content were expressed. Left-wing content was even lower, appearing only once after expressing right-wing preferences. Centrist posts were relatively high, particularly after right-wing preferences were indicated. No hate speech was observed, though one piece of hostile content appeared later in the journey. Country specific themed posts dominated the type of political posts seen throughout the user journey, where other themes included other history and misogyny.

Figure 44: User journey illustration (Romania, TikTok, Right-wing avatar)

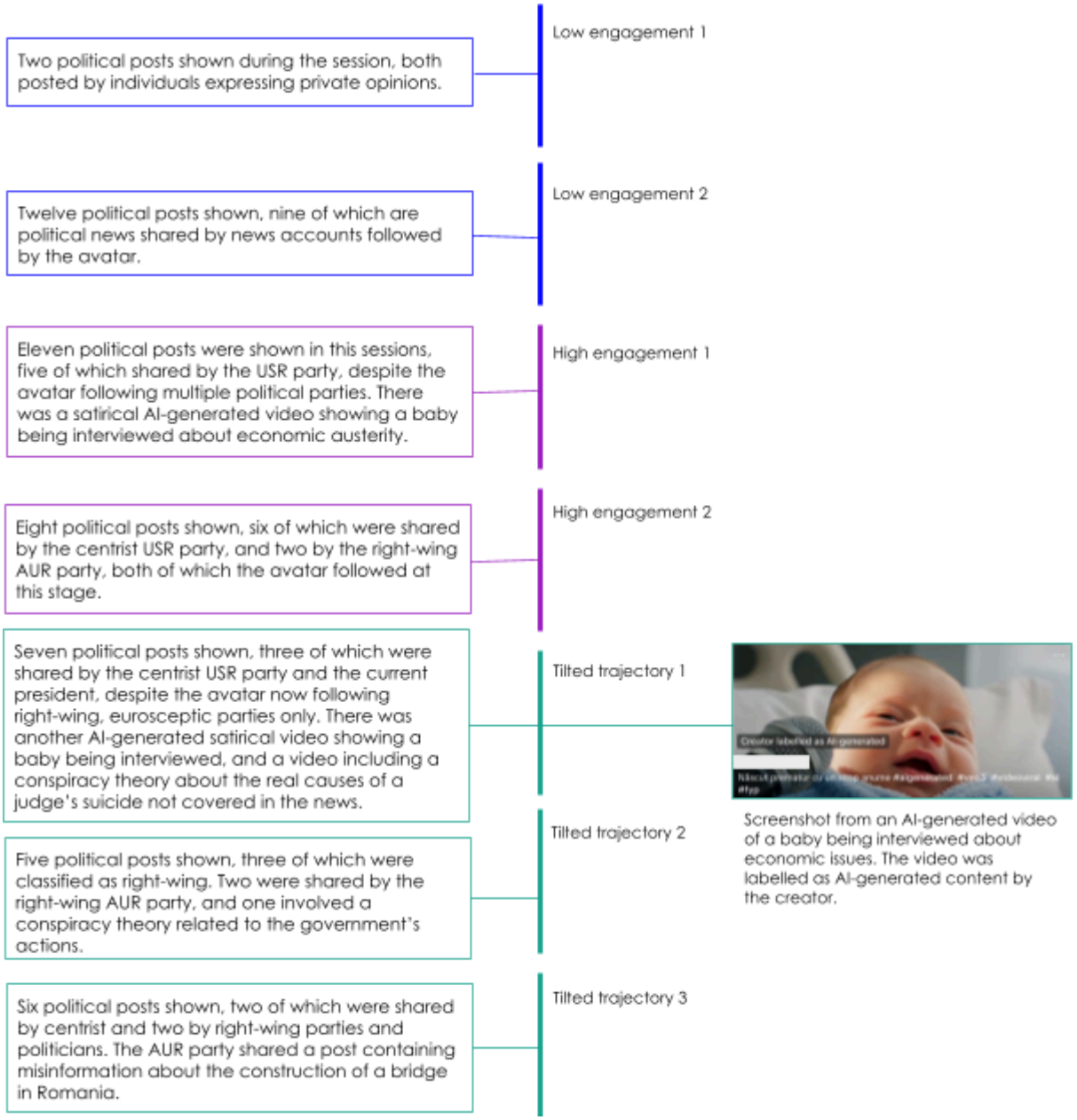
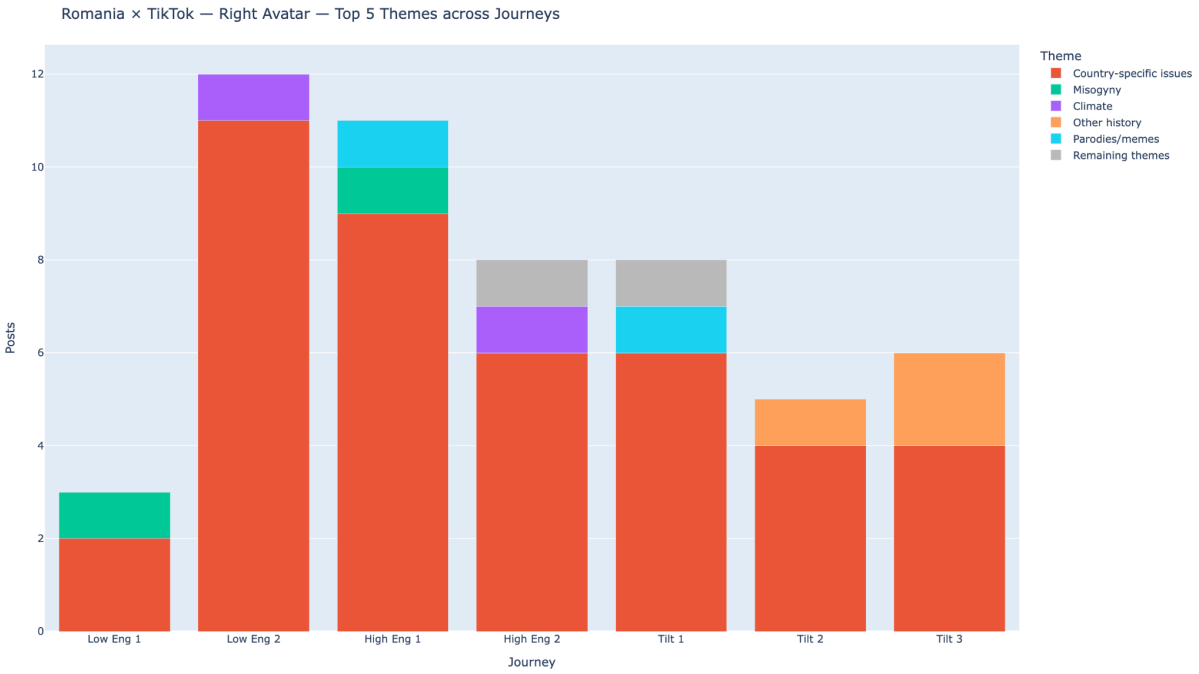


Figure 45: Thematic analysis of political posts encountered in user journey (Romania, TikTok, Right-wing avatar)



Left-wing avatar

On this user journey (see Figures 46 and 47), the avatar did not encounter conspiracy theories or misinformation, though malinformation was observed. The avatar encountered low levels of left-wing and right-wing content on this journey. No hate speech or hostile content was observed. Country-specific political posts were a key theme throughout this user journey. Other themes included misogyny, the climate and the remaining variety of themes.

Figure 46: User journey illustration (Romania, TikTok, Left-wing avatar)

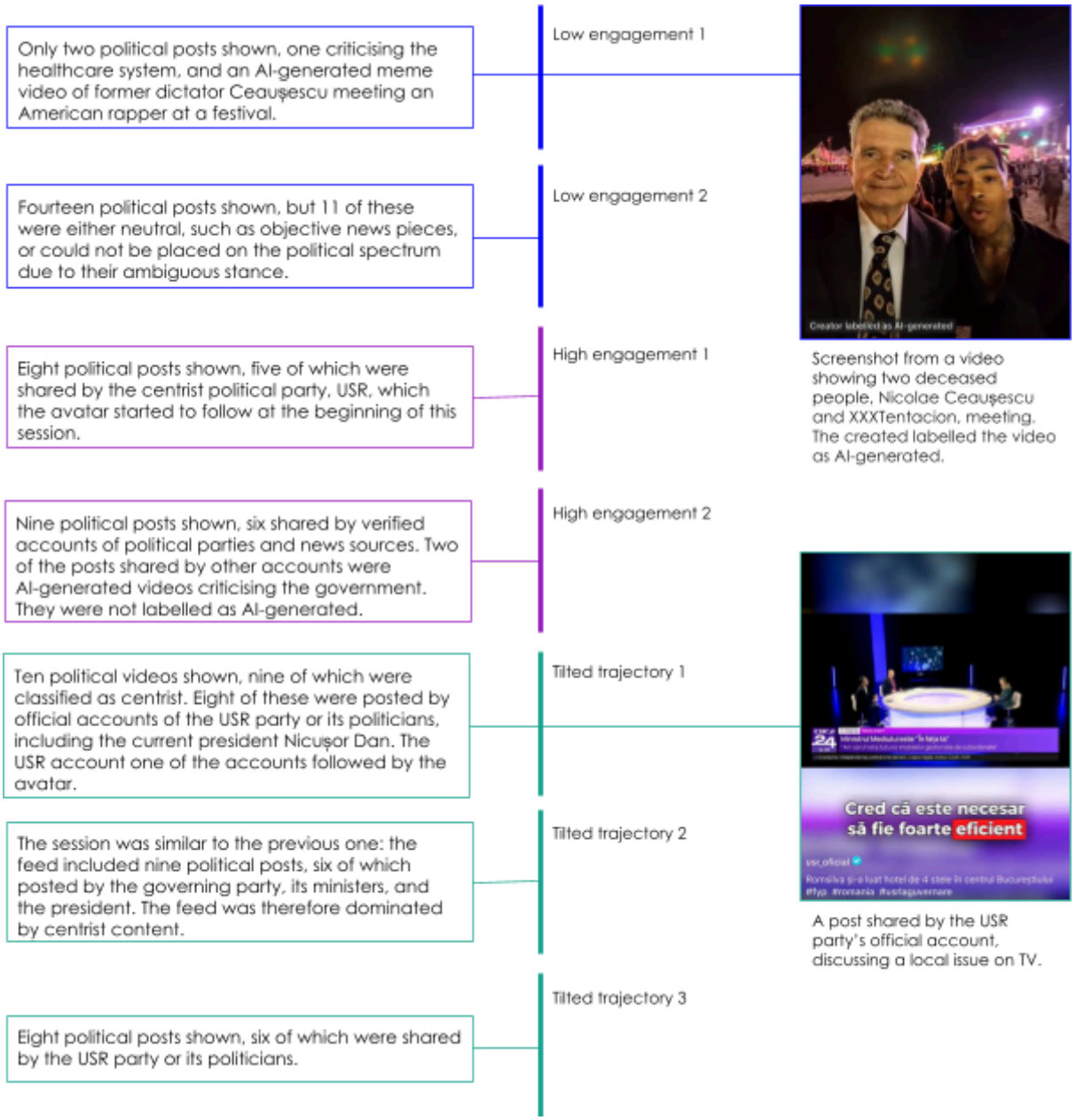
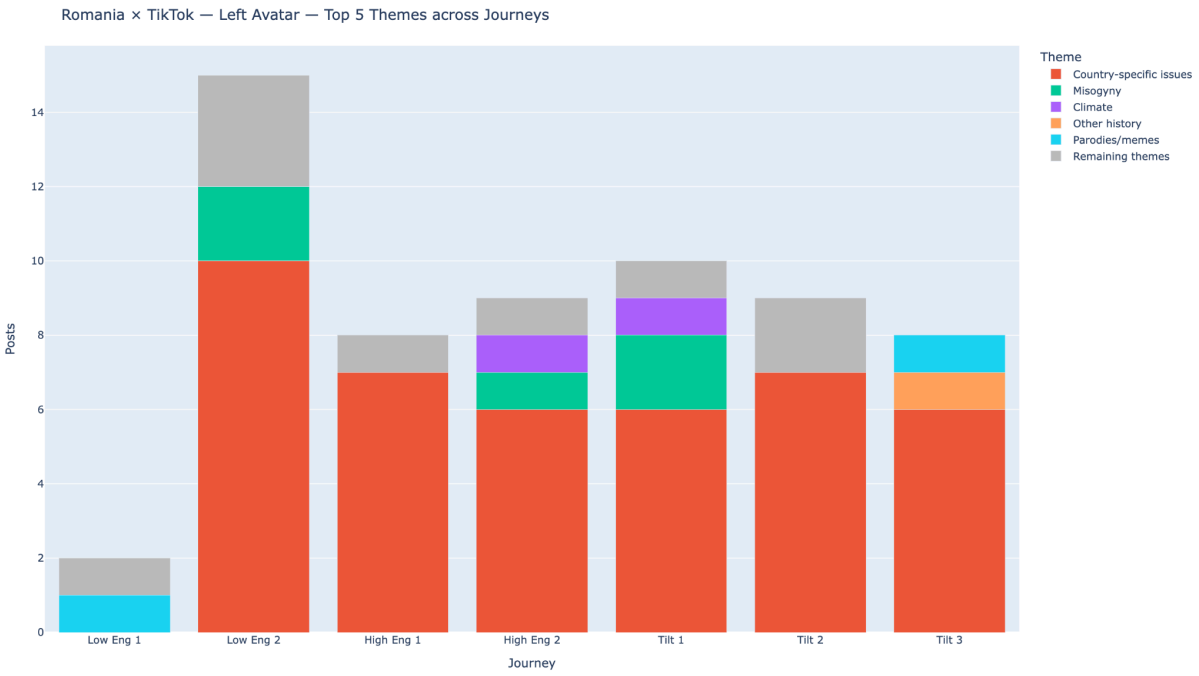


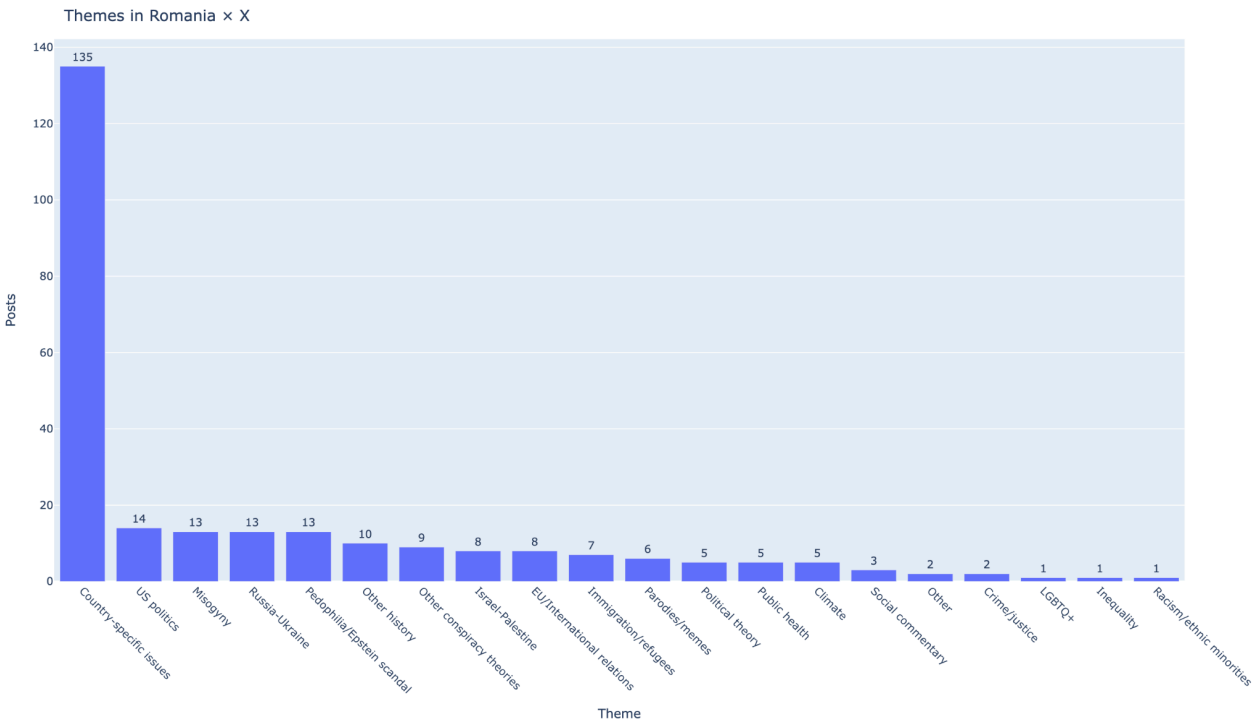
Figure 47I : Thematic analysis of political posts encountered in user journey (Romania, TikTok, Left-wing avatar)



X

The political content encountered by the avatars on X in Romania was broadly balanced across left- and right-wing posts, with centrist material appearing at similar levels. Levels of misinformation, malinformation, and conspiracy theories were generally low, though the left-wing avatar encountered higher amounts early in the browsing sessions. One instance of hate speech and some hostile content were observed for the right-wing avatar, while the left-wing avatar experienced only a single instance of hostile content. Political posts were frequent at the start of the journeys, with recurring themes including country-specific issues, US politics, misogyny, pedophilia-related topics and the Russia–Ukraine war (see Figure 48 for the thematic breakdown of posts). Overall, the types and volumes of political content remained relatively stable over time, and the tilt expressed by the avatars appeared to have minimal effect on the content displayed.

Figure 48: Thematic analysis of all political posts (Romania, TikTok)



Right-wing avatar

On this user journey (see Figures 49 and 50), the avatar encountered a moderate level of mis- and malinformation, and conspiracy theories, with one instance of hate speech (religious extremism) and some hostile content. Political posts were frequent, peaking at initial engagement, with right-wing and left-wing content highest when the avatar first interacted with the platform. Of the political posts, a prominent theme was country-specific issues; other themes included pedophilia/the Epstein scandal, misogyny and US politics.

Figure 49: User journey illustration (Romania, X, Right-wing avatar)

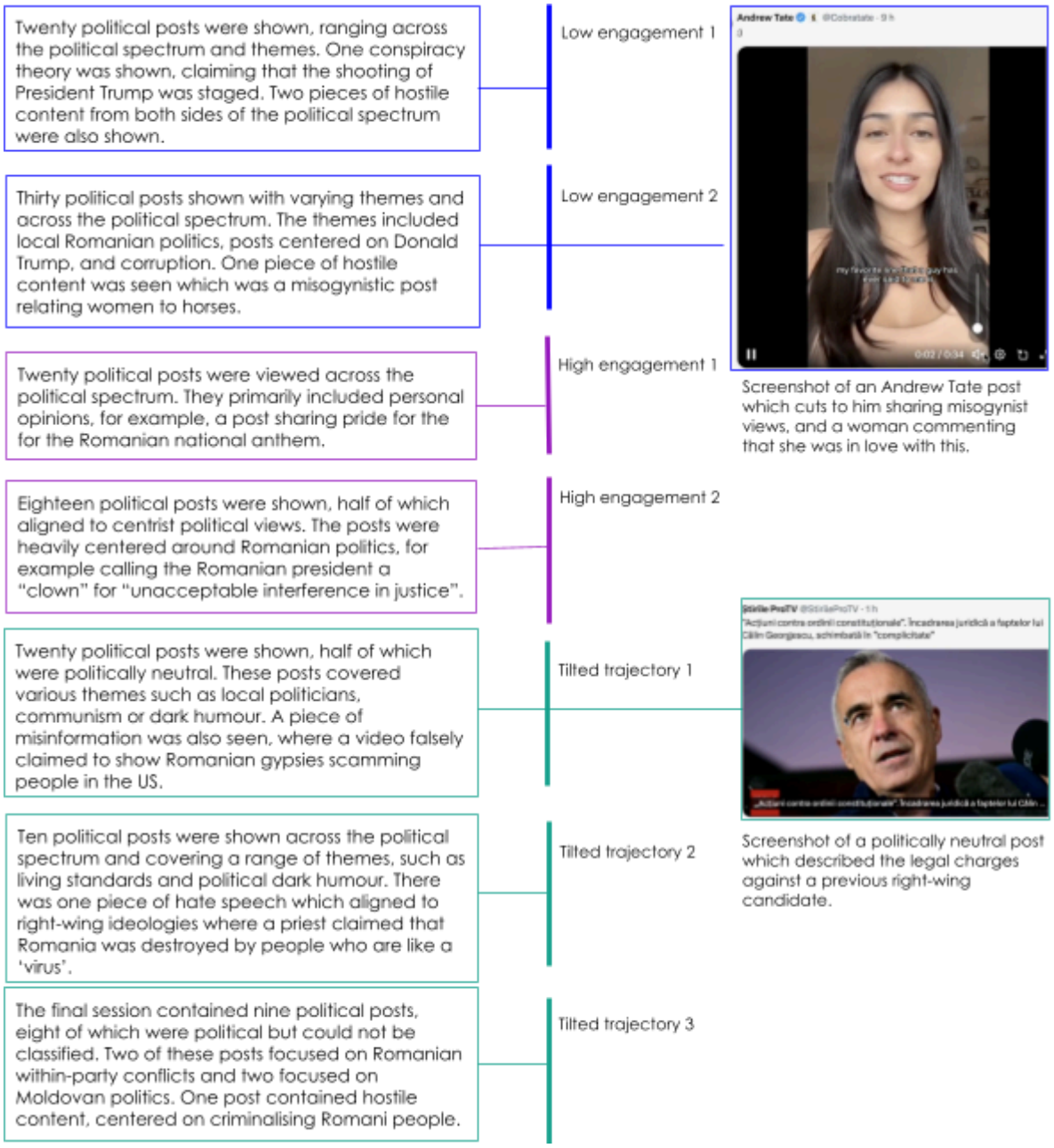
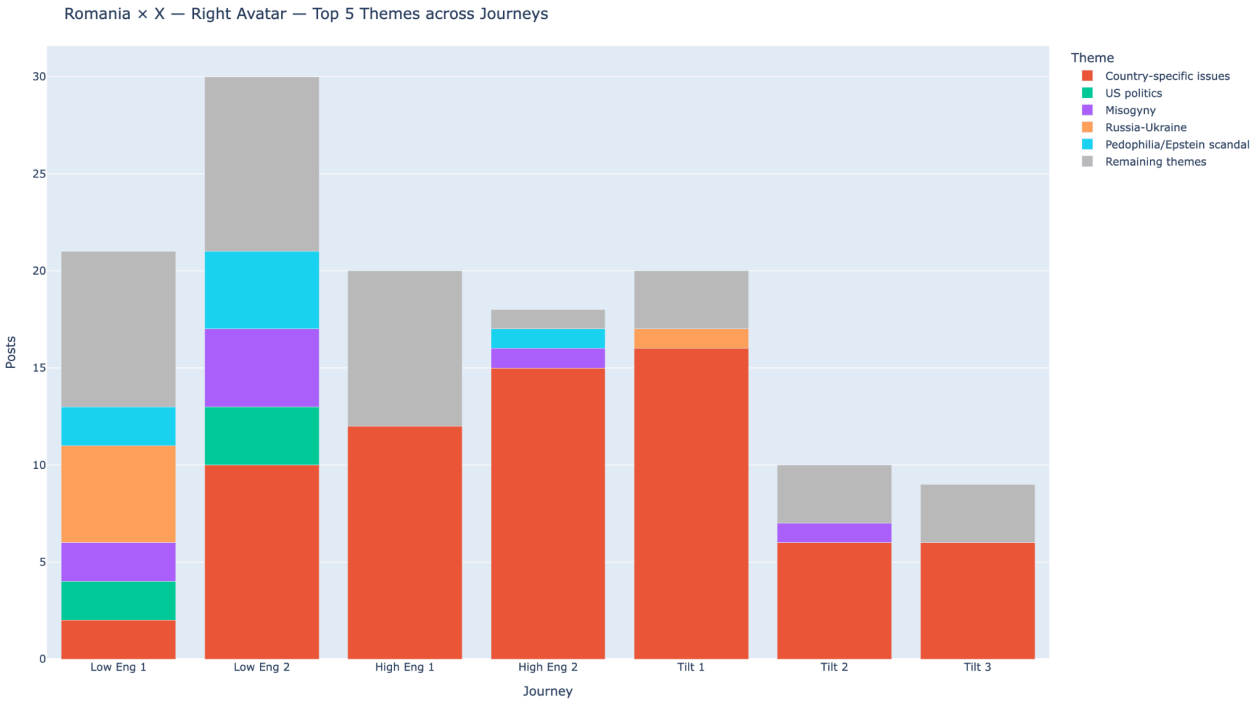


Figure 50: Thematic analysis of political posts encountered in user journey (Romania, X, Right-wing avatar)

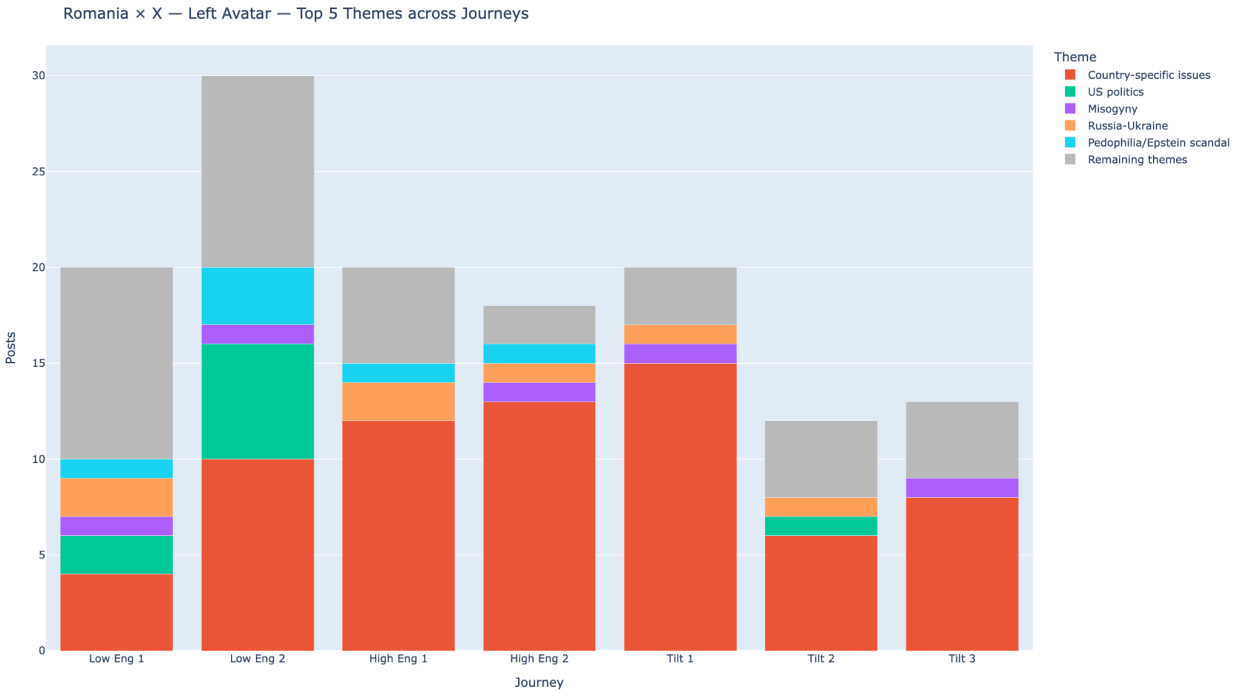


Left-wing avatar

On this user journey (see Figures 51 and 52), the avatar initially encountered a high level of conspiracy theories and misinformation, along with some malinformation, but no hate speech and only one instance of hostile content. Political posts were frequent at first, with right-wing content particularly prominent before decreasing as the avatar expressed left-wing preferences. However, once this was expressed left-wing posts also decreased. This suggests the algorithm reacted to user signals in this case. Centrist content appeared at consistent levels throughout. On this user journey, country-specific issues featured heavily; other post themes included US politics, the Russia-Ukraine war and misogyny. This avatar saw a diverse variety of themes, with some focus on historic politics and wars.

Figure 51: User journey illustration (Romania, X, Left-wing avatar)

Figure 52: Thematic analysis of political posts encountered in user journey (Romania, X, Left-wing avatar)



Audit 2 (Finland)

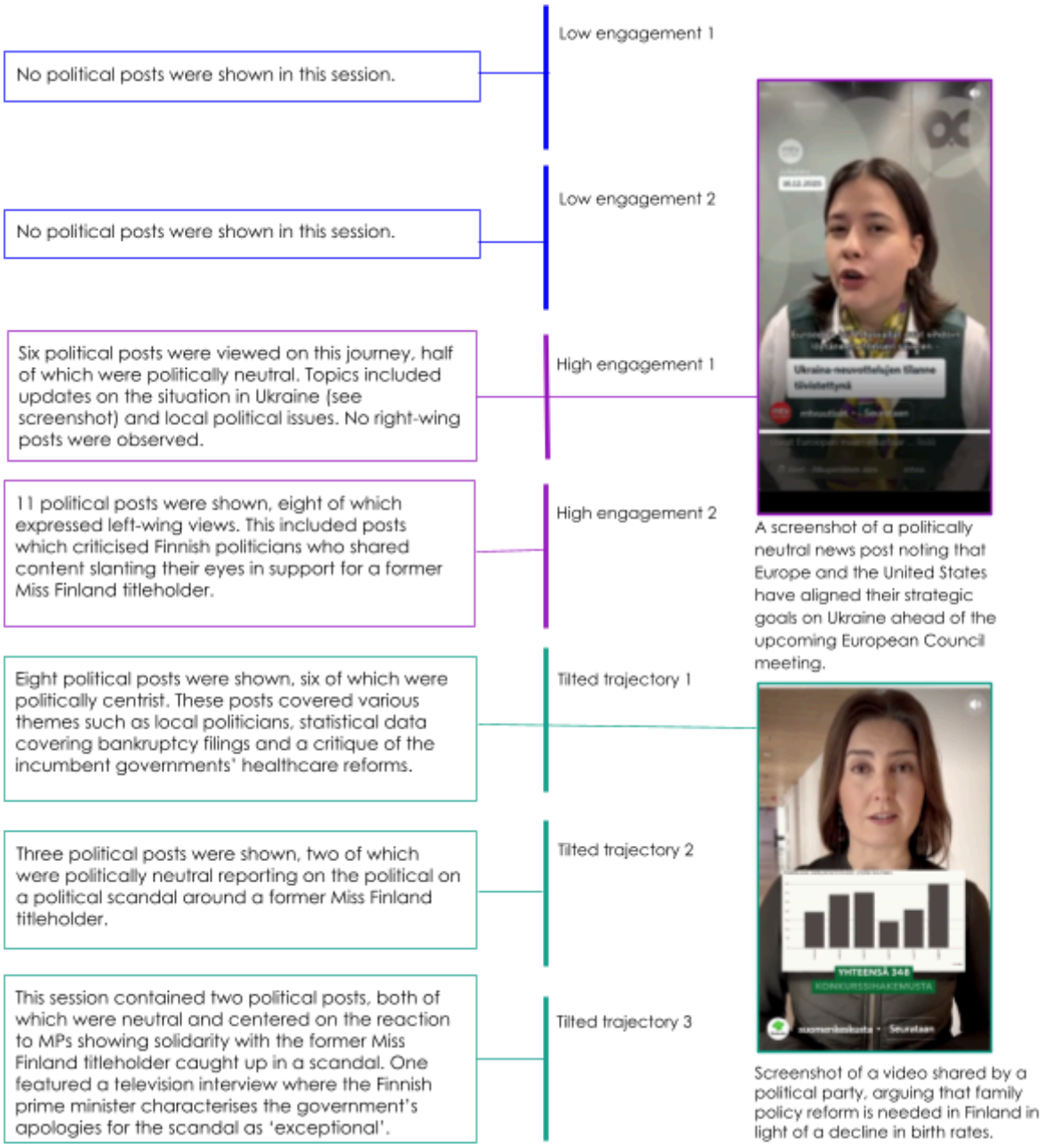
Instagram

The two journeys on Instagram were similar: both started without any political content, and once the avatars expressed political interest, the platform showed a moderate amount of political posts. Both avatars saw neutral news and centrist or left-of-centre opinion pieces, irrespective of which parties they followed or liked the content of. This reveals a left-wing bias on Instagram in the second audit, and highlights that simple actions, like following political parties, had no influence on the political content recommended. None of the content shown was problematic, i.e. the avatars did not observe mis- or malinformation, hate or hostile speech, or conspiracy theories. None of the political posts were identified as AI-generated either.

Right-wing avatar

Initially, no political content was shown to the avatar. Once the avatar indicated interest in politics, some political news and left-of-centre political content appeared on the feed. Throughout the 'High-Engagement' and 'Tilted Trajectory' phases, the avatar mostly saw neutral news and left-of-centre or centrist political content. No right-wing content was shown, despite the avatars' signalled interest in right-of-centre politics. The most dominant theme was the scandal involving the former Miss Finland titleholder and political reactions to the video she posted slanting her eyes, allegedly mocking Chinese people. No problematic content was observed in this journey.

Figure 53: User journey illustration (Finland, Instagram, Right-wing avatar)

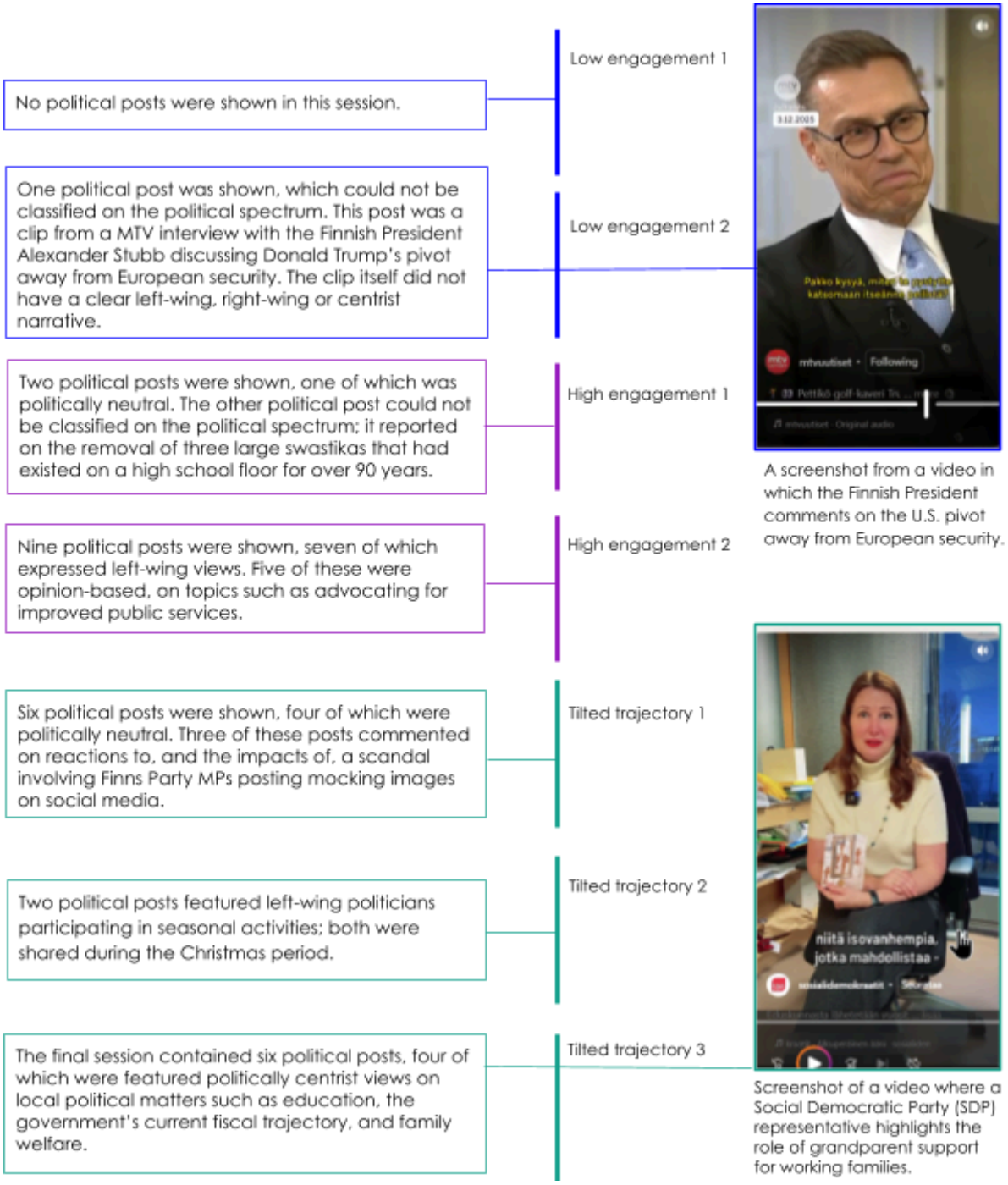


Left-wing avatar

In the 'Low-Engagement' phase of this journey, only a single political post was shown. Once the avatar indicated interest in politics, each browsing session had a moderate amount of political content, ranging from two to nine such posts per session. These were all neutral news or centrist or left-of-centre opinions, aligning with the avatar's stated preferences. No problematic content was shown. Some of the

political content shown was influenced by current events, such as reflections on an ongoing scandal related to the former Miss Finland titleholder and Christmas greetings shared by politicians.

Figure 54: User journey illustration (Finland, Instagram, Left-wing avatar)



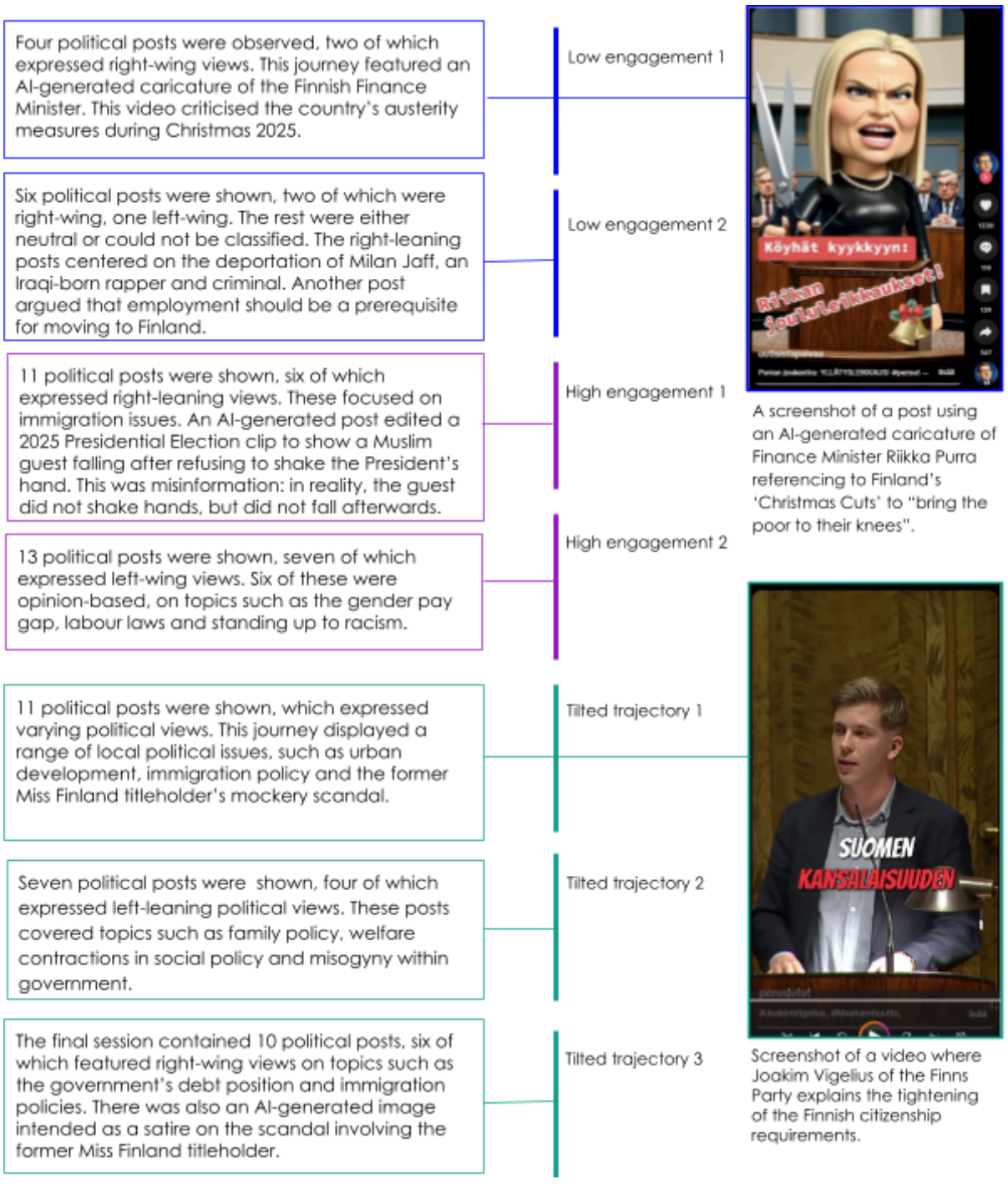
TikTok

The journeys on TikTok contained a moderate level of political content, fluctuating between one and 13 posts per browsing session. Overall, there was no clear left- or right-wing bias. The posts focused mostly on domestic issues, especially topical news pieces like the scandal involving the former Miss Finland titleholder and political reactions to this. However, TikTok feeds showed some AI-generated videos, which focused on racial mockery. They either recreated — and in some cases altered — divisive scenes from the news, or created stereotypical ethnic minority figures for parodistic purposes. While some of these videos were simply offensive or had an anti-immigrant sentiment, others contained hostile speech too.

Right-wing avatar

Throughout this journey, the avatar saw moderate to high levels of political content, ranging between four and thirteen political posts per session. In some sessions, right-of-centre content was more dominant than centrist or left-of-centre content, but in others, left-of-centre content was more frequent. This suggests that the avatar's signalled preference for right-wing content was not consistently reflected in the content recommendations. Notably, this journey included several AI-generated posts meant as political satire. One of these mocked the Finance Minister from a leftist stance, one mocked and spread misinformation about a Muslim woman refusing to shake the hand of the President at an event, and one made fun of the scandal involving the former Miss Finland titleholder.

Figure 55: User journey illustration (Finland, TikTok, Right-wing avatar)



Left-wing avatar

Initially, the avatar saw a low amount of political content, which increased after following and engaging with the content of political parties. The posts seen covered a broad range of political views, without an obvious bias. However, there were several AI-generated posts observed, typically commenting on racial or ethnic issues. Some reflected on the scandal involving the former Miss Finland titleholder, and some contained hostile speech directed against Romani people.

Figure 56: User journey illustration (Finland, TikTok, Left-wing avatar)

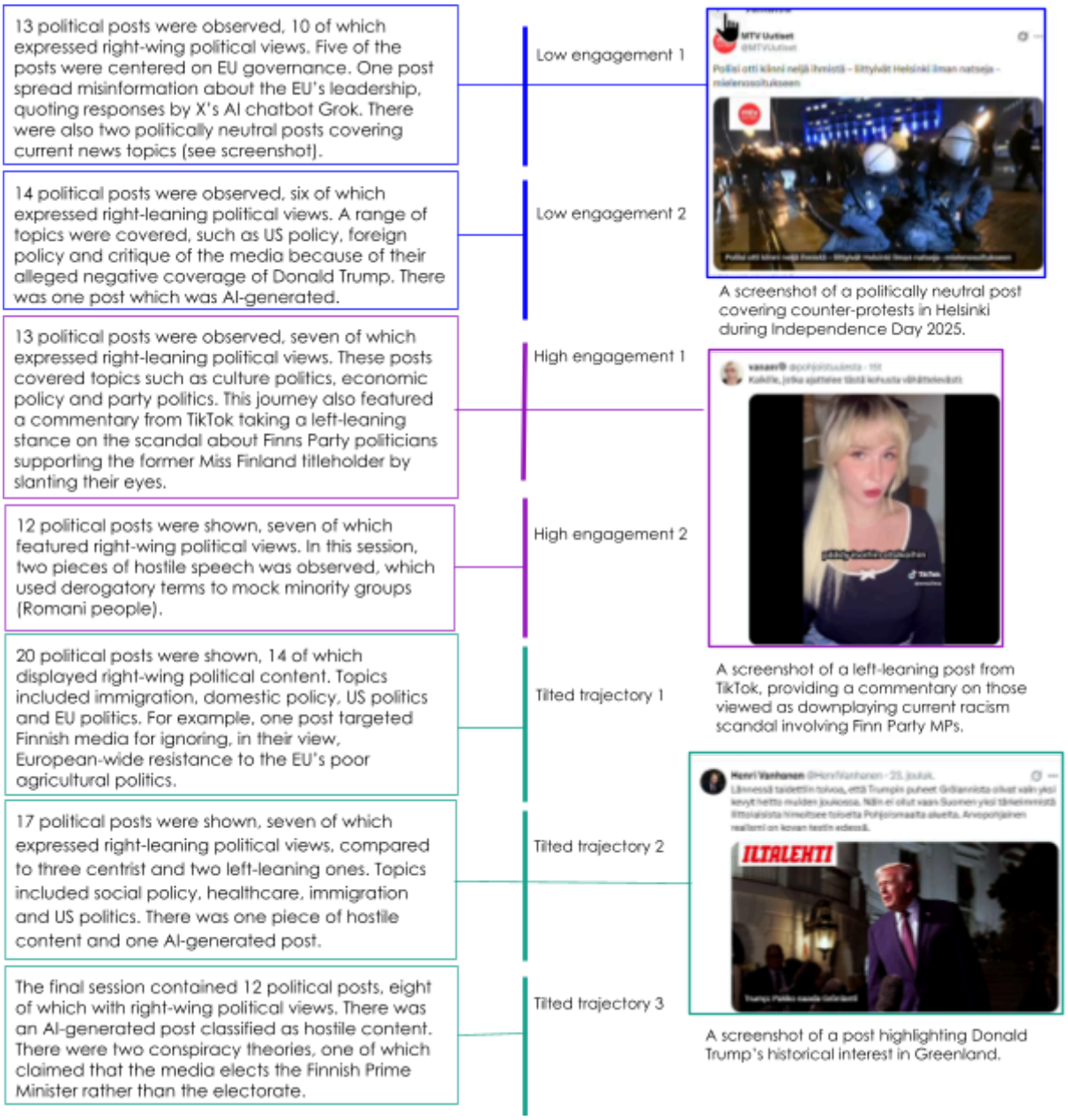
X

Both journeys on X had a high level of political content, covering Finnish and international (mostly USA-related) issues. The majority of the browsing sessions were dominated by right-wing content. The signalled interests of the avatar did not seem to influence recommendations: both avatars saw high levels of political content even before indicating political interest, and signalling interest in left-wing politics did not lead to more left-wing content being shown. The avatars saw problematic content across their journeys, most notable AI-generated content including hostile speech. These posts typically targeted ethnic minorities, e.g. Romani and Black people. They also saw misinformation and conspiracy theories related to domestic and EU politics.

Right-wing avatar

This journey had a consistently high level of political content with a right-wing bias. Regardless of the signalled interest of the avatar, X recommended between 12 and 20 political posts per browsing session. In all the sessions, right-wing content was dominant, but there were some centrist, left-wing or neutral political posts shown too. The posts focused on both domestic and international politics, with a notable presence of USA politics. The avatar saw a range of problematic content as well, including misinformation and conspiracy theories about the EU and Finnish politics, and hostile speech against Romani people.

Figure 57: User journey illustration (Finland, X, Right-wing avatar)



Left-wing avatar

This journey was characterised by a high level of political content, regardless of how much political interest the avatar showed. In the 'Low-Engagement' phase, the avatar saw a roughly balanced mix of right-wing, left-wing and centrist content, covering both domestic and international issues. However, after entering the 'High-Engagement' phase, the political content became dominated by right-wing

views. The avatar saw multiple posts containing hostile speech, some of them AI-generated. These focused on ethnic minorities and racial issues.

Figure 58: User journey illustration (Finland, X, Left-wing avatar)