Scrutinizing Systemic Risks in Personalized Recommender Systems Through Sock-Puppet Auditing of VLOPs

LUKA BEKAVAC, University of St. Gallen, Switzerland
JANNIS STRECKER, University of St. Gallen, Switzerland
KIMBERLY GARCIA, University of St. Gallen, Switzerland
SIMON MAYER, University of St. Gallen, Switzerland
AURELIA TAMÒ-LARRIEUX, University of Lausanne, Switzerland

Very Large Online Platforms (VLOPs) use personalized recommender systems to optimize their main performance metric: attention-based user engagement. In doing so, these systems might however amplify systemic risks by promoting controversial or polarizing content, thereby exacerbating issues such as misinformation, societal polarization, and the manipulation of civic discourse. To mitigate these risks, regulations such as the European Union's Digital Services Act (DSA) mandate increased data access and transparency, including for the auditing of personalized recommender systems. However, the data access provided by VLOPs remains limited—often restricted to specific user demographics, aggregate statistics, or curated datasets—hindering meaningful oversight. Consequently, new methods are needed to audit recommender systems effectively at the user level. In this paper, based on an analysis of the legal context and technical alternatives for data access, we present SOAP, the System for Observing and Analyzing Posts. SOAP is an open-source framework for auditing recommender systems using sock-puppet accounts. It enables fine-grained user-level analysis beyond the constrained data access typically provided by platforms. We detail SOAP's technical implementation and evaluate its ability to scrutinize systemic risks. Additionally, we tested SOAP in a workshop with over 100 participants and observed a measurable increase in participants' algorithmic literacy. This demonstrates SOAP's potential not only for research and regulatory auditing, but also as an educational framework to foster public awareness of algorithmic influence.

CCS Concepts: • **Information systems** → **Personalization**; **Content ranking**; *Recommender systems*; • **Human-centered computing** → **Social media**; • **Applied computing** → **Law**; • **Social and professional topics** → **Technology audits**.

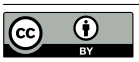
Additional Key Words and Phrases: Platform Regulation, Social Media, Black-Box testing, Systemic Risks, Filter Bubbles, DSA, Sock-Puppet Auditing

ACM Reference Format:
Luka Bekavac, Jannis Strecker, Kimberly Garcia, Simon Mayer, and Aurelia Tamò-Larrieux. 2026. Scrutinizing Systemic Risks in Personalized Recommender Systems Through Sock-Puppet Auditing of VLOPs. *ACM Trans. Recomm. Syst.* 1, 1 (January 2026), 48 pages. https://doi.org/10.1145/3795516

1 Introduction

Social media platforms—especially dominant Very Large Online Platforms (VLOPs)—play a central role in shaping public discourse, influencing access to information, and structuring societal

Authors' Contact Information: Luka Bekavac, luka.bekavac@unisg.ch, University of St. Gallen, St.Gallen, Switzerland; Jannis Strecker, jannis.strecker-bischoff@unisg.ch, University of St. Gallen, St.Gallen, Switzerland; Kimberly Garcia, kimberly.garcia@unisg.ch, University of St. Gallen, St.Gallen, Switzerland; Simon Mayer, simon.mayer@unisg.ch, University of St. Gallen, St.Gallen, Switzerland; Aurelia Tamò-Larrieux, aurelia.tamo-larrieux@unil.ch, University of Lausanne, Lausanne, Switzerland.



ACM 2770-6699/2026/1-ART
https://doi.org/10.1145/3795516

interactions [98]. Their recommendation systems act as algorithmic gatekeepers, determining what users see and when. This algorithmic mediation introduces *systemic risks*, as defined by the EU's Digital Services Act (DSA) [66], which include the dissemination of illegal content, negative impacts on fundamental rights, threats to electoral integrity and civic discourse, and harms to public health or well-being. Systemic risks, unlike isolated incidents, refer to platform-induced patterns that are wide-reaching, reproducible, and potentially destabilizing to societal or democratic structures [58, 67, 108]. These risks are not merely theoretical. Real-world incidents—such as cases of election interference [44], the dissemination of illegal content [51], and negative effects on fundamental rights [42]—have gained extensive media coverage and triggered formal investigations by the European Commission against platforms like TikTok and X. The European Commission has addressed dominant online platforms directly through formal proceedings[1] and official letters[2]. Growing societal unease is also prompting individual action, as evidenced by the partial exodus of public figures and institutions from the platform X in late 2024, reflecting mounting concern over the structural influence of VLOPs on democratic resilience[3]. Closely tied to these systemic risks is the widespread integration of *personalization* into social media platforms, fundamentally reshaping how content and advertisements are delivered to users [95]. While personalized recommender systems leverage user behavior data to enhance engagement and incentivize further disclosure of personal information [16], they also introduce substantial challenges and risks. High-profile incidents—such as the selective, intransparent suppression of user content (*shadowbanning* [54]), data misappropriation exemplified by the Cambridge Analytica scandal [100], and ethically controversial experiments such as Facebook's emotional contagion study [60]—underscore the seriousness of these risks. Such systemic risks are further exacerbated by phenomena like *filter bubbles*, where personalization primarily confines users to content that reinforces their existing views and interests [85]. This narrowing of information sources restricts users' worldviews [11], amplifies misinformation, and may intensify societal polarization [46]. These developments highlight the urgent need for greater oversight and for establishing guidelines and standards that ensure that recommender systems serve the public interest rather than undermine it [63].

In response to these concerns, researchers and policymakers are demanding increased transparency from social media platforms, in particular with respect to their data collection, processing, and profiling practices. However, while scholars, civil society, and regulatory bodies argue that open data access is essential to investigate systemic risks [12, 24, 78], the so-called *APIcalypse* has marked a significant retreat from this ideal, as many platforms now impose stringent limits on previously available data streams [12]. This includes: (a) contractual constraints (e.g., within Terms of Service agreements) that explicitly prohibit data scraping and other independent collection methods; and (b) technical countermeasures, such as blocking specific tools used for automated data collection, and restricting access to previously open APIs [64]. Regulators have already responded to this development: the European Union introduced the Digital Services Act (DSA) [36], a landmark regulation that seeks to enhance platform accountability and transparency.

However, despite new regulatory instruments on research APIs and transparency instruments already having entered into force, data access remains limited in practice [56]. The research APIs mandated under the DSA—intended to facilitate independent scrutiny—often fail to meet

[1]The European Commission opened formal proceedings against TikTok in relation to systemic risks: https://ec.europa.eu/commission/presscorner/detail/en/ip_24_926 and https://ec.europa.eu/commission/presscorner/detail/en/ip_24_6487. Last accessed January 13, 2026.

[2]Thierry Breton emphasized the societal responsibility of the X platform, noting that "[w]ith great audience comes greater responsibility," see https://x.com/ThierryBreton/status/1823033048109367549. Last accessed January 13, 2026.

[3]See e.g., https://www.theguardian.com/media/2024/dec/11/from-x-to-bluesky-why-are-people-abandoning-twitter-digital-town-square. Last accessed January 13, 2026.

researchers' needs and hinder the intended comprehensive auditing because of discretionary platform interpretations of what constitutes "publicly available" information that are coupled with technically restrictive environments (i.e., data enclaves) and delayed implementation of data-sharing provisions [10, 94, 105]. Overcoming these shortfalls requires alternative approaches that do not require privileged platform cooperation: data scraping, black-box testing, and sock-puppet auditing[4] stand out as critical methodologies for auditing recommender systems [14, 70]. While these methods have faced resistance through legal and technical barriers [64], they remain vital for investigating whether and how platform operations influence civic discourse, fundamental rights, and the overall health of democracies [17]. Importantly, however, *technical* auditing solutions for recommender systems need to be designed in a way that closely aligns with the *legal* context—this is the gap that we address with this article and that expands on the recently proposed *System for Observing and Analyzing Posts* (SOAP).[5]

In this article, we make three main contributions: First, based on an examination of the current regulatory background—in particular under the DSA—we structure the data landscape, show what data is required for auditing, introduce existing data access approaches, discuss alternatives, and analyze their benefits and limitations. Second, we show that sock-puppet auditing is a viable methodology to overcome current challenges. We give a detailed account of SOAP, a techno-legal framework designed to complement the DSA's data-access provisions by leveraging legally and ethically valid and technically feasible data scraping and black-box testing methodologies even in the absence of privileged platform cooperation. Third, we present and discuss results from a workshop with over 100 participants that used SOAP, demonstrating that SOAP significantly improved their Algorithmic Media Content Awareness (AMCA [127]). To aid future audits of VLOPs, SOAP is open-source and publicly available[6]. Our contributions show that novel techno-legal approaches are required to address VLOPs' systemic risks, and that in particular SOAP's sock-puppet auditing significantly empowers researchers as well as civil society to conduct meaningful, independent investigations into platform practices. We show that this at the same time can increase platform transparency and user awareness with respect to personalized recommender systems employed by VLOPs. SOAP as a tool—and our approach as a methodology—thereby represent a recommended way forward to support the DSA's overarching ambition: to foster a transparent, accountable, and more equitable digital environment.

2 Background and Related Work

The DSA defines systemic risks only in broad terms [81], leaving considerable room for interpretation. Under Article 34(1) [36], the DSA divides systemic risks into four main categories [55]: the dissemination of illegal content, negative effects on fundamental rights, threats to civic discourse and electoral processes, and harms related to issues such as public health or mental well-being. In the following, we focus on personalized recommender systems as a critical vector through which systemic risks may be exacerbated. Specifically, we argue that personalization mechanisms—by reinforcing phenomena such as filter bubbles—can intensify the impact of systemic risks. We use the phenomenon of filter bubbles as an analytical lens to illustrate the scope and conditions under which risks to civic discourse and electoral processes can emerge (Section 2.1). We then discuss how the platforms can be technically audited in today's regulatory environment (Section 2.2),

[4]Sock-puppet auditing [7, 96] involves deploying automated bots that mimic real user behavior to analyze how platforms personalize content and influence user engagement.

[5]This article is an extended version of a paper entitled "From Walls to Windows: Creating Transparency to Understand Filter Bubbles in Social Media" [10], which has been presented at the second NORMalize Workshop at the 18th ACM Conference on Recommender Systems (RecSys'24).

[6]https://github.com/Interactions-HSG/SOAP

introduce options to access VLOP data through means recommended by the DSA and through other alternatives (Section 2.3), and explore automated deductive coding as an efficient mean to analyze large amounts of VLOP data (Section 2.4).

2.1 Systemic Risks Stemming from Personalized Recommender Systems

Recent election years have heightened concerns about systemic risks related to *negative effects on civic discourse, electoral processes, and public security*. In particular, personalization on online platforms has attracted attention for narrowing the range of information available to users, thereby distorting public discourse and aggravating these risks. Such narrowing complicates efforts by campaigns, media, and policymakers to effectively engage diverse audiences. Scholars from social sciences [13, 31] and computer science [49, 73, 116] have extensively studied these dynamics, focusing particularly on phenomena such as *societal polarization* [8, 48, 102] and *filter bubbles* [29, 62, 85]. Filter bubbles, as first conceptualized by Eli Pariser in 2011 [85], describe personalized digital environments, such as social media feeds, where algorithms lead to intellectual isolation and social fragmentation. Pariser's concept since has faced criticism for lacking empirical specificity, leading to inconsistent findings across studies [47, 72, 114]. To systematically study filter bubbles, Michiels et al. more recently proposed a measurable and empirically grounded definition: "A technological filter bubble is a decrease in the diversity of a user's recommendations over time, in any dimension of diversity, resulting from the choices made by different recommendation stakeholders" [72, p.275]. Such a technological filter bubble thus emphasizes four elements:

- *Diversity:* A filter bubble manifests itself as a reduction in the structural, topic, or viewpoint diversity of recommendations. In which:
  - *Structural diversity* is the variety of information sources or suppliers, such as news outlets, users, or brands.
  - *Topic diversity* refers to the breadth of subjects or topics presented, such as sports, politics, or science.
  - *Viewpoint diversity* highlights the range of perspectives or stances on a given topic, including opposing or contrasting viewpoints.
- *Recommendations:* The phenomenon is observed as the personalized and curated content for users.
- *Time:* Filter bubbles emerge gradually as systems learn more about user preferences.
- *Recommendation stakeholders:* Multiple actors, including users and platforms, influence the recommendation process.

Even though, filter bubbles are often studied as a problematic phenomenon, reduced content diversity might be desired [29], e.g., in the context of *protective filter bubbles* as "an algorithmically curated information ecosystem that shields users from threats to psychological and physical safety, including targeted threats such as hate speech, discrimination, and political persecution and generalized threats such as distressing media" [31, p.2]. Importantly, such dynamics do not emerge solely from algorithmic design. Prior work emphasizes the role of users' self-selection of preferred content [29], which interacts with non-explicit, inferred feedback signals used for personalization. Together, these mechanisms can shape—and in some cases intentionally narrow—the diversity of content in personalized feeds without necessarily constituting a harmful or externally imposed filter bubble. However, decreased diversity (e.g., in social media feeds) may also exacerbate polarization, hinder democratic processes, and amplify systemic risks such as misinformation, manipulation, and discrimination [46]. For instance, TikTok's algorithm amplifies polarizing and hard-to-verify content about the Israel-Hamas conflict [99], which could be linked to risks such as public health and mental well-being. In another example, TikTok's algorithm created a "bespoke reality" for Kamala Harris

supporters during the 2024 U.S. presidential election, amplifying campaign narratives while isolating users from alternative viewpoints [97]. Measuring systemic risks, such as societal polarization or the erosion of democratic discourse, remains challenging due to the opaque nature of platform algorithms, as the systemic risk can stem from the "functioning" and "use" of a platform [55]. Scholars propose various approaches to address these risks, focusing on conceptual and empirical strategies. Calabrese et al. [17] emphasize the need for systematic methodologies to assess content diversity and detect filter bubbles or echo chambers that distort public discourse. They propose analyzing the spread of polarized content and conducting real-time studies during critical events, such as elections, to observe the amplification of divisive narratives. Marsh [68] highlights the difficulty of accessing comprehensive algorithmic data, such as ranking metrics and system logs, under the DSA. To address this, Marsh suggests relying on indirect methods like black-box testing and integrating evidence collection with platform-provided transparency tools. Loi [66] proposes evaluating systemic risks through normative frameworks, such as measuring media pluralism and false positives in content moderation. Recent work also organizes algorithm-audit methodologies into four complementary study designs—risk-uncovering, reverse-engineering, interface-design, and risk-measuring—framed as distinct objectives within the DSA risk-management cycle [82]. This distinction helps clarify what kinds of evidence different audits can generate, and motivates tools that operationalize repeatable study designs under constrained data-access conditions. However, access to critical algorithmic data, such as ranking metrics or personalization factors, is often restricted by platforms, creating a significant barrier for researchers. This limitation underscores the necessity of alternative methodologies. In response, researchers have begun to move beyond broad definitional work, advocating for more systematic, data-driven methodologies to identify, analyze, and mitigate systemic risks. This shift underscores a growing recognition that without actionable frameworks and concrete tools, the complexities of VLOPs' influence on public discourse and civic life will remain elusive.

2.2 Emerging Legal Tools to Audit Systemic Risks

The DSA encourages scrutiny and oversight of VLOPs in regards to systemic risks [55]. This includes the facilitation of external audits by vetted researchers and the implementation of transparency measures, such as publicly accessible databases and access to publicly accessible data [70]. Together, these initiatives aim to open VLOPs' "black boxes", enabling a better understanding of how algorithmic processes—specifically recommender systems—foster or mitigate systemic risks [70].

There are two closely related methods to enable these independent investigations: First, *publicly available data* may be collected through *data scraping* methods. Second, *sock-puppet audits* [96] enable *black-box testing* and allow researchers to simulate user interactions, gather data on platform outputs under controlled conditions [115], and draw conclusions about underlying recommendation dynamics [116]. Here, third-party auditors can rely on the DSA to develop and refine black-box auditing strategies based on legally and publicly accessible data [70].

Scraping involves programmatically retrieving publicly available data—such as search results or recommendation feeds—without relying on platform-provided APIs or internal access [70]. This data collection method serves as a critical enabler for black-box testing, which treats the platform as an opaque system. *Black-box testing*, examines the platform's behavior by analyzing the relationships between inputs (e.g., user interactions or queries) and outputs (e.g., recommended content or search results) across varied scenarios [14]. Earlier studies have successfully employed this approach to quantify the prevalence of misinformation and analyze content recommendations on platforms [1, 6, 52, 84, 91, 107].

Sock-puppet audits involve creating artificial user accounts or "virtual agents" that emulate genuine user behavior to probe how recommender systems respond to different content-consumption

patterns [70, 96, 115]. By precisely controlling variables such as viewed content, follow patterns, or search queries, auditors can detect biases, manipulative tendencies, or harmful recommendation cascades. An advantage of sock-puppet audits is their ability to implement experimental research designs by varying profile characteristics analogous to (field) experiments in social science [115]. All elements of the "puppet" are held constant except for information in the user profile.

Recent systematic reviews of algorithm audits help contextualize these approaches and clarify where sock-puppet designs are particularly valuable. Bandy [7], synthesizing 62 audit studies and adopting the taxonomy of Sandvig et al. [96], distinguishes five recurring audit families: *direct scraping*, *crowdsourcing*, *sock-puppets*, *carrier-puppets*, and *code audits*. Across the reviewed literature, direct scraping is most common and well suited for descriptive characterization, but it often lacks randomization or manipulation. Sock-puppet and crowdsourced designs are less frequent, yet enable more controlled comparisons by systematically varying inputs and user characteristics. The review also highlights recurring constraints for sock-puppet audits, including platform restrictions and challenges in approximating realistic behavior, and notes that audit attention is concentrated among a small number of organizations and platforms. Urman et al. [118] extend this line of work with a systematic review of 176 audits and argue that, under contemporary behavior-based personalization, the most consequential distinction is often whether data are collected under *non-personalized* versus *personalized* conditions. Because browser automation is increasingly used even for direct scraping and non-personalized data collection, the mere use of automated agents is not, by itself, informative about research design. This motivates an explicit focus on how personas are specified and how user behavior is operationalized in personalized audits. Taken together, these reviews underscore both the continued value of sockpuppet audits and the practical barriers to conducting them at scale on highly personalized VLOPs.

By combining sock-puppet auditing and scraping, researchers can conduct *black-box testing* and generate empirical evidence about platform recommendations [91]. For instance, researchers can use "blank" sock-puppets initially and gradually "train" them by exposing them to specific content sets [10]. Parallel scraping of platform search results and recommendation feeds, along with metadata gleaned from publicly available channels, may then help triangulate findings and highlight patterns [107]. Prior research has demonstrated the utility of sock-puppeting approaches and scraping in revealing platform behaviors. Haroon et al. [49] used over 100'000 sock-puppets to audit YouTube's recommendation system, identifying how ideological content is amplified through watch trails, with evidence of bias toward extreme and problematic content. Similarly, Ledwich et al. [62] explored the concept of radical bubbles, finding that while YouTube often recommends similar content, evidence for deliberate radicalization pipelines was limited. Building on these methods, Tjaden et al. [115] applied sock-puppet audits to TikTok's recommendation system during Germany's 2024 regional elections, revealing significant biases: politically neutral users were 3–4 times more likely to encounter right-wing populist content than mainstream party content. These studies highlight the effectiveness of scraping and sock-puppet audits in identifying systemic risks, providing valuable empirical evidence for independent oversight.

2.3 Platform Data Access Options

A basic prerequisite to audit VLOPs is the access to platform data, a cornerstone of the DSA [36, 55]. By mandating data access provisions, the DSA seeks to empower researchers to scrutinize platform operations, foster accountability and support the development of robust systemic risk mitigation measures [70]. In theory, this approach should provide the transparency necessary for academia and civil society to investigate the societal impacts of VLOPs. In practice, however, the realization of the DSA's vision faces significant challenges. While the law requires platforms to share data on their design, functioning, and mitigation measures, researchers often encounter practical barriers [56] as

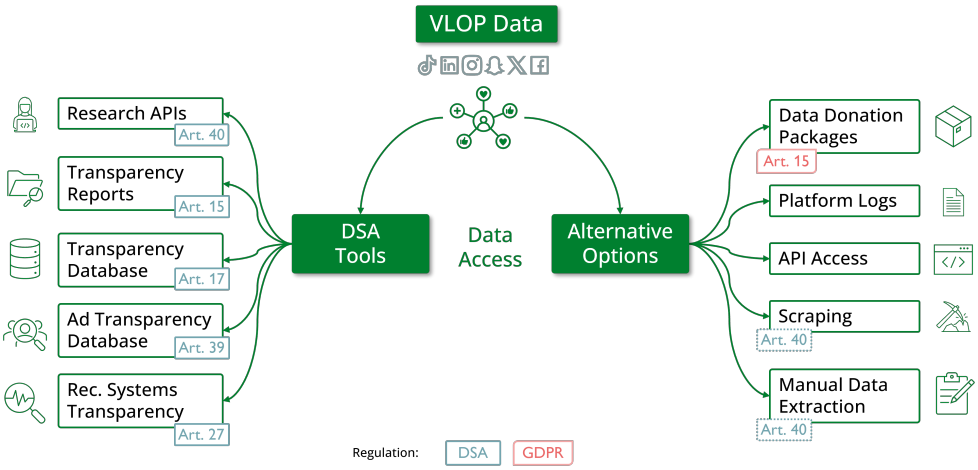


Fig. 1. Overview of VLOP data access options using DSA tools (left) or alternative options (right).

data access often depends on platform cooperation [70] and discretionary data availability [25]. These barriers additionally include restrictive interpretations of "publicly accessible" data [78], infrastructural and technical limitations on data retrieval [94], and difficulties in verifying whether the data provided is sufficient and accurate for independent investigations [40]. Such challenges hinder the comprehensive scrutiny required to address systemic risks stemming from recommender systems effectively, and leave researchers with limited insights into platform operations and systemic risk factors.

2.3.1 Platform Data Access Options provided by the DSA. To facilitate the scrutiny of VLOPs' recommender systems and to analyze the challenges associated with their use (see Figure 1 on the left), the DSA mandates the following transparency and data access methods.

Research APIs. Under Article 40(12) of the DSA, researchers should be able to access data that is already publicly available. Platforms may operationalize this obligation through a range of mechanisms, including granting permission for scraping, providing data on an ad hoc or request-based basis, or offering dedicated Research APIs [41]. In practice, however, there are numerous structural and technical shortcomings, as well as divergent interpretations of what constitutes "publicly accessible" [56]. Although data access under Article 40(12) is not limited to Research APIs, these interfaces have increasingly become the primary means through which very large online platforms formalize and mediate researcher access [41]. The absence of centralized and clearly defined procedural guidelines has given rise to a fragmented landscape in which each VLOP effectively crafts its own version of compliance, thereby restricting meaningful research and inhibiting comprehensive platform audits [25]. For instance, some Platforms argue that certain proposed studies do not sufficiently contribute to "the detection, identification, and understanding of systemic risks," thereby justifying data withholding under the guise of DSA compliance [56]. Furthermore, the definition of "publicly accessible information" differs from platform to platform, e.g. Meta (as of December 2024) excludes profiles with fewer than 25,000 followers and TikTok does not provide data originating from users under the age of 18. Additionally, the scope of the available data points is further restricted by the platforms, e.g., TikTok's Research API provides access to only a small fraction of the metadata parameters observable through other data access methods [94]. Beyond definitional ambiguities, the technical systems set up for public data access

place researchers into highly controlled virtual environments. Meta's Virtual Data Enclave (VDE)[7] and TikTok's Virtual Compute Environment (VCE)[8], for instance, restrict researchers strongly, e.g., by preventing the use of external libraries, or by requiring that all imports, exports, and even intermediate results (including aggregated statistics and notes) must be pre-approved by the platforms. In the absence of cohesive enforcement and explicit technical standards, Article 40(12) of the DSA devolves into a patchwork of highly conditional, platform-dictated terms [25]. Hence, researchers find themselves navigating a complex maze of shifting eligibility criteria, controlled technical infrastructures, and minimized data outputs [25]. Such limitations curb researchers' ability to conduct independent, large-scale, and reproducible analyses of platform behaviors. By confining publicly accessible data and its analysis to heavily monitored virtual enclaves, platforms can effectively shape the boundaries of inquiry. This dynamic gives VLOPs disproportionate control over what questions are even answerable [78], narrowing the space for methodological innovation and inhibiting the development of robust, evidence-based understandings of platform-level risks.

Transparency Reports. Article 15 of the DSA mandates that VLOPs publish periodic transparency reports, providing insight into their content moderation practices. These reports must cover content moderation measures, including the volume of content taken down, responses to orders from national authorities, and the performance metrics of automated moderation systems. Given the influence and reach of VLOPs, these reports also need to include information about content moderation teams, specifically their qualifications and linguistic skills, to demonstrate their readiness to manage diverse content responsibly. These transparency reports are meant to be a crucial basis for scrutinizing platform operations, as they provide researchers and regulators with initial data to evaluate compliance with the DSA's objectives. However, current practices rely heavily on the self-reported data of VLOPs. This highlights the importance of external audits, such as those enabled under Article 40 data access of the DSA, to independently assess the claims made in these transparency reports and to ensure accountability. Even though, the European Commission maintains a list of transparency reports submitted by platforms[9] (as noted on the Commission's website) not all platforms retain previous reports online, complicating continuous public access to historical data on platform practices.

Transparency Database. The DSA requires VLOPs to enhance transparency regarding their content moderation decisions. Specifically, Article 17 mandates that providers of hosting services must provide a clear and specific statement of reasons to any affected recipients when they impose restrictions based on illegal content or incompatibility with their terms and conditions. Additionally, Article 24(4) obliges these platforms to upload these statements to the publicly accessible *DSA Transparency Database*[10]. However, the official platform provided by the DSA leaves room for improvement, and shortcomings have been highlighted by researchers. Kaushal et al. [57] argue that while there are some transparency gains, compliance remains problematic due to the database's structure allowing platforms considerable discretion in their transparency practices. For example, platforms overwhelmingly choose to remove content based on their Terms of Service (ToS)—accounting for 99.8% of removals—rather than on the basis of illegal content (0.2%). Similarly, Trujillo et al. [117] found that (i) platforms adhered only partially to the database's intended philosophy and structure; (ii) the database structure is partially inadequate for platforms' reporting needs; (iii) platforms exhibited substantial differences in their moderation actions; (iv) a significant fraction of the data is inconsistent; and (v) the platform X (formerly Twitter) presents

[7] https://transparency.meta.com/researchtools/meta-content-library. Last accessed January 13, 2026.
[8] https://developers.tiktok.com/doc/vce-getting-started. Last accessed January 13, 2026.
[9] https://digital-strategy.ec.europa.eu/en/policies/dsa-brings-transparency. Last accessed January 13, 2026.
[10] https://transparency.dsa.ec.europa.eu/. Last accessed January 13, 2026.

the most inconsistencies. While these issues suggest that the current design of the Transparency Database falls short of its goal to enhance accountability and transparency in some areas of content moderation practices, it can serve as a crucial starting point for external scrutiny under Article 40(4) of the DSA, which allows researchers to request specific data access. Transparency reports and databases thus provide a preliminary benchmark for verifying the accuracy of the data disclosed by platforms. Demonstrating the necessity of data access requests through these resources can hold platforms accountable.

Advertisement Libraries. Article 39 requires VLOPs to compile and make publicly available a repository containing detailed advertisement information [36]. This repository should include, e.g., the content and presentation duration of the advertisement, the entity responsible for the advertisement, the targeted population, and the number of recipients the ad was presented to. According to a report by the Mozilla Foundation, none of the surveyed advertisement libraries provided by Platforms were fully operational as of April 2024 [21], nor did these libraries provide researchers and civil society groups with the tools and data needed to effectively monitor the impact of VLOPs' advertisements on potential upcoming elections.

Recommender System Transparency. The DSA imposes specific obligations on platforms to enhance transparency regarding their recommender systems. According to Article 27(1) of the DSA, platforms are required to explain in their terms and conditions, using clear and intelligible language, the main parameters used by their recommender systems, as well as any options available to users to modify or influence those parameters. This provision aims to ensure that users are appropriately informed about how recommender systems impact the presentation of information and how they can influence the content displayed to them [90]. Furthermore, Article 27(2) specifies that the explanation must include the most significant criteria determining the recommendations, such as the content and its ranking, along with the reasons for their relative importance. Article 27(3) requires platforms to provide a directly and easily accessible functionality that allows users to modify their preferred options related to the recommender system parameters. In addition, Article 38 mandates that VLOPs that use recommender systems must provide at least one option for each recommender system that is not based on profiling. This requirement was also upheld by the Amsterdam District Court in October 2025, forcing Meta to make the option of non-profiled feed directly and easily accessible for its users[11]. Currently, platforms have primarily focused on explaining how recommendations are generated and delivered, with varying levels of detail, rather than implementing accessible functionalities that allow users to modify the recommendation parameters [90]. For instance, Instagram[12] and TikTok[13] provide explanations, in their respective help centers, on why content is served to users. Even though, platforms like Instagram offer some user controls, such as the ability to indicate reasons for disliking content or filtering out certain hashtags, there is limited functionality for users to directly intervene in the parameters of the recommender systems. In a best-case scenario, compliance with Articles 27 and 38 of the DSA would require VLOPs to implement effective functionalities that allow users to influence the algorithmic parameters and alter the output of recommendations [90]. This could include providing various levels of control features and explanations of different complexity to cater to the diverse skills of users. For instance, users might be able to adjust the degree of personalization, select the types of content they wish to see or avoid, and choose which of their data can be used for profiling-based recommendations [90].

[11] https://the-platform-law.com/2025/10/09/the-bits-of-freedom-ruling-the-first-step-in-private-dsa-enforcement/
[12] https://transparency.meta.com/features/explaining-ranking/ig-explore/. Last accessed January 13, 2026.
[13] https://support.tiktok.com/en/using-tiktok/exploring-videos/how-tiktok-recommends-content. Last accessed January 13, 2026.

2.3.2 Alternative Platform Data Access Options. While the DSA establishes new mechanisms for data access and scrutiny, these measures often prove insufficient in practice; as platforms continue to restrict their APIs and limit researcher access, the academic and civil society communities have sought alternative methods to investigate systemic risks and platform behaviors without relying solely on platform-provided access modalities or cooperation [12]. Figure 1 on the right provides an overview of five primary approaches currently available for data access that do not rely on DSA provisions, namely: Data Donation Packages, Platform Logs, API Access, Scraping, and Manual Data Extraction. Each method offers distinct opportunities to analyze systemic risks by focusing on content diversity, user exposure, and the dynamics of personalized recommender systems. Each method offers unique advantages but is constrained by technical and/or legal challenges.

Data Donation Packages. In the European Union, Article 15 of the General Data Protection Regulation (GDPR) [34] mandates that individuals have the right to access all personal data that a platform has stored about them. Such Data Donation Packages (DDPs) provide a new and promising method to collect timestamped and content-based information about social media use [119]. This approach, however, is constrained by low compliance rates, delays in receiving data packages, and the complexity of extracting information from these packages [119]. Different social media platforms present varied and complex data formats, and issues related to privacy compliance add additional layers of complexity [119]. Data rights under the GDPR have the advantage of being legally enforceable and enabling access to very fine-grained data. However, they also raise a number of legal, ethical, and methodological issues whose significance varies depending on the specific research projects [4]. For example, previous research has utilized data download packages from users for social media analysis [120, 126]. However, since the data and social media feeds are deeply personal spaces that reflect individual values and preferences, high standards of privacy and data protection must be maintained [39].

Platform Logs. Platform logs involve capturing data from web browsers, apps, or devices via plugins or applications, which are automatically transferred to a research server [24]. Unlike data donations, this method does not require users to actively share their data or engage in any interactions. Such tools include, e.g., browser extensions like "Who Targets Me"[14] which aims to make online advertisements more transparent, or Data Selfie[15] which used to enable Facebook users to see their own data traces. These tools passively collect data on user interactions and the content shown to users. Research utilizing platform logs has investigated areas such as misinformation and algorithmic bias [79], or adolescent wellbeing [112]. While platform logs can provide valuable insights into user interactions and content delivery mechanisms, using them for research might be complex, since social media companies often restrict or disable tools that collect such data, citing violations of their terms of service or legal concerns [23].

API Access. Access to APIs for the public after the so-called *APIcalypse* [12] (see Section 1) has become significantly restricted. The public APIs of social media platforms are either non-existent for researchers or are so heavily restricted that they threaten the reproducibility and replicability of Social Media research [24]. For instance, Twitter/X's API has undergone substantial changes that limit access for non-commercial users. The free tier now allows only minimal API requests, providing access to limited endpoints that are insufficient for comprehensive data collection. To obtain more extensive data, Twitter/X offers enterprise tiers starting at $500,000 per year for access to just 0.3% of the company's tweets. Researchers argue that this cost is prohibitively high for the amount of data provided, making it inaccessible for most academic research projects [109]. Other

[14] https://whotargets.me/de/. Last accessed January 13, 2026.
[15] https://dataselfie.it. Last accessed January 13, 2026.

examples include the shutdown of research tools using platform data, such as Netvizz in 2019[16], and limitations placed on CrowdTangle[17], where no new users were allowed from 2022, and a complete shutdown occurred in 2024.

Scraping. Scraping involves using tools designed to gather data from websites for data analysis, either through automated scripts or explicit programming. These methods do not use an official API and have been used in the past, but have often been banned or restricted by the platforms [64]. Additionally, the practice of web scraping is often legally complex [24] and currently exists in a legal grey zone [64], with varying opinions from different authorities, including data protection agencies [28]. Responding to this uncertainty, e.g., the Institute for Strategic Dialogue calls for regulators to recognize the value of mixed-methods approaches such as diverse data collection methods used by researchers to understand the broader implications of social media platforms on individuals and society [78]. So far, the legal complexity of scraping has also led public interest researchers to limit transparency regarding how they scrape and conduct their investigations using platform data. For instance, the project "Auditing TikTok" [94] states on their code website: "In these repository you'll find some of the code we've developed to look into TikTok. We can't publish everything to not risk our analysis, but if you are interested in learning more, please get in contact". These legal and ethical considerations are more thoroughly addressed in the design principles of SOAP (see Section 3.4).

Manual Data Extraction. Manual data extraction involves collecting publicly available data, such as post content, images, and engagement metrics, directly from social media platforms [24]. While insights from manual methods have been valuable—as demonstrated by Amnesty International's report on TikTok's promotion of self-harm content [53] and studies like Recommending Toxicity [5] and Safer Scrolling [87], which uncovered the amplification of misogynistic and toxic content—these approaches face significant drawbacks. For example, researchers in these projects manually interacted with and analyzed content over extended periods, using archetypes to guide user behavior. Although these efforts revealed algorithmic biases, the labor-intensive process limits the number of posts or users analyzed, resulting in small, non-representative sample sizes; highlighting the need for more scalable, automated approaches to studying platform dynamics.

2.4 Deductive Coding of Social Media Data

Once social media data has been collected, it must be analyzed to yield meaningful insights into potential systemic risks. This typically requires identifying recurring themes, content categories, or normative signals across large volumes of multimodal data. Given the scale and diversity of platform content, deductive coding becomes a critical step in structuring and interpreting datasets in a consistent and reproducible way. Effective coding approaches must therefore balance scalability with contextual sensitivity, particularly when analyzing politically charged or culturally nuanced material. Social media data presents numerous methodological challenges [111]. First and foremost is the *volume*: platforms generate massive streams of content, much of which is short-form and multimodal in nature. For example, personalized feeds often consist of reels or stories ranging from a few seconds to one minute, combining text, visuals, and audio. Prior research highlights the extensive resource demands of manual coding methods, which can require hundreds of person-hours [116] and is practically infeasible at scale [116]. The sheer volume of textual, visual, and audio data thus forces researchers to adopt shortcuts to manage this workload. For instance,

[16] http://thepoliticsofsystems.net/. Last accessed January 13, 2026.
[17] https://apnews.com/article/meta-crowdtangle-research-misinformation-shutdown-facebook-977ece074b99adddb4887bf719f2112a. Last accessed January 13, 2026.

researchers frequently analyze only transcripts of videos or use video snippets to reduce the data volume [37, 122]. However, selecting only a portion of the data from individual social media channels for analysis carries the risk of either losing relevant information, over-representation, or gathering data that is of little value.

Second, social media content is inherently *heterogeneous* and context-dependent [101, 111]. Posts may contain slang, abbreviations, hashtags, images, memes, deepfakes, or stitched audio. They often remix content from multiple sources, sometimes across languages, and with little regard for conventional grammar [80]. Such variability renders manual or rule-based coding methods brittle and time-consuming. To address these challenges, previous studies have explored the use of Large Language Models (LLMs) to support deductive coding tasks [19, 20, 43, 123]. LLMs—particularly multimodal models—are well suited for this purpose: they can process diverse inputs (text, image, audio) and scale to large datasets, offering a viable path toward consistent and efficient annotation. In practice, this makes LLM-based coding especially useful for scalable descriptive categorization, hypothesis generation, and comparative analyses across sock-puppets and conditions (e.g., tracking shifts in topical exposure over time). It should be acknowledged that LLMs are not without limitations. However, the goal is not to replace human judgment, but to triage, summarize, and structure large-scale content for downstream analysis and interpretation. In Section 3.3, we detail our approach to validating coding reliability and reflect on where automation succeeds—and where human oversight remains critical.

2.4.1 Limitations of Deductive Coding with Multimodal-LLMs. LLMs can lack the nuanced understanding and contextual judgment that human coders bring to complex social data, which may impact the depth and authenticity of the analysis. Moreover, concerns about the epistemic limitations of LLMs, including biases in training data and misinterpretation of social and cultural cues, underscore the need for careful evaluation of their outputs. Essentially, the same limitations that apply to standard LLMs also come with multimodal LLMs. Below, we discuss the limitations encountered when using the multimodal LLM in SOAP and how we mitigated them.

Ambiguous and Subjective Concepts. Open-ended questions and not clearly defined concepts are difficult to measure. For example, letting an LLM code what is potentially *woke* or *insulting* leaves too much room open for interpretation. Also, studies have shown the political preferences of LLMs [89, 93], which may influence how LLMs code on political topics and coding tasks. One such domain is conspiracy theories and misinformation. Due to the nature of conspiracy theories, it is challenging to create primer prompts or use an LLM to effectively flag such content [37]. Furthermore, as LLMs have a knowledge cutoff, very recent events and news cannot be sufficiently fact-checked. Hence, implementing cross-validation techniques to compare machine and human coding for reliability and accuracy before using a primer prompt for deductive coding can be employed for intra and inter reliability testing (see Section 3.3).

Harmful Content. One problem encountered in flagging extremist content is the safety filters and limitations of the models, such as those from Google Gemini Models.[18] Most publicly available LLMs assess content against a list of safety attributes, which include harmful categories and topics that can be considered sensitive. For each attribute, a model assigns one safety score based on the probability of the content being unsafe and another safety score based on the severity of harmful content. Content that exceeds a certain threshold is blocked, and the LLM does not provides any response. This poses significant challenges for tasks like extremist content labeling, as the models

[18] https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/configure-safety-attributes#vertex-ai-gemini-api. Last accessed January 13, 2026.

may refuse to process or return results for such content. To overcome this, one might use self-hosted LLMs without safety filters or infer meaning from the safety filter responses themselves. For example, the safety score indicating violent content as seen in Appendix A could serve as an indicator even if an LLM response is blocked.

Inherent Biases. As covered by extensive literature [65, 113], LLMs have biases based on the data they have been trained on and their underlying architecture. This must be taken into account when using automated coding instead of human coding, especially when coding topics or domains that may be underrepresented on an LLM [124].

3 Measuring Systemic Risks: SOAP «the System for Observing and Analyzing Posts»

As outlined in Section 2.3, the current landscape of data access for investigating systemic risks on VLOPs is fraught with challenges. Official data access mechanisms under the DSA often fall short, providing limited data that hinders comprehensive investigations. Alternative methodologies, such as data scraping and black-box testing, offer potential solutions but are constrained by significant legal and technical barriers. Together, these issues underscore a pressing need for tools that can operate within the legal framework of the DSA while overcoming the technical and legal limitations of existing methods. To address these challenges and further advance the study of systemic risks, we developed the *System for Observing and Analyzing Posts* (SOAP) [10], a novel tool designed to collect and analyze data from VLOPs.[19] SOAP is specifically aimed at studying systemic risks through sock-puppet auditing. The system allows for the automated exploration and navigation of content and auditing of the recommender systems on VLOPs. While measuring systemic risks of dominant online platforms fulfills an important societal need, it must be done in adherence to legal and ethical standards. Therefore, during the development of SOAP, we took into account and carefully weighed different competing interests during its design phase (see 3.4 ; this required the integration of legal and technical expertise in our research team. Hence, we present SOAP as a *techno-legal tool for investigating systemic risks.*

3.1 SOAP Overview

SOAP is a framework for auditing personalized recommender systems on VLOPs. It leverages configurable sock-puppet accounts to simulate user behavior, measure content exposure, and evaluate how recommendation systems contribute to systemic risks—such as filter bubbles, political radicalization, or harmful content amplification. Building on prior sock-puppet audit designs and the methodological gaps highlighted in recent reviews (Section 2.2), SOAP provides (i) automated longitudinal capture via platform-facing private interfaces that logs full post-level metadata, (ii) fine-grained, configurable personas and interaction policies beyond keyword- or hashtag-based steering, (iii) multimodal LLM-based deductive coding with an accompanying reliability-testing procedure, and (iv) techno-legal design choices intended to keep the workflow compatible with prevailing regulatory advancements on data access and platform auditing. SOAP is based on two key innovations. First, it introduces *active puppets* [115], which extend traditional sock-puppet methods by dynamically adapting their behavior during runtime. These puppets are guided by a "primer prompt," which defines their topical interest (e.g., climate change denial or election misinformation) and steers their actions accordingly. This allows researchers to model and test a wide range of user personas and behavioral patterns. Second, SOAP integrates multimodal large language models (LLMs) to analyze the diverse and voluminous content typical of modern social platforms. Rather than relying on manual coding alone, SOAP automates the deductive coding process, identifying

[19]The code is publicly available: https://github.com/Interactions-HSG/SOAP.

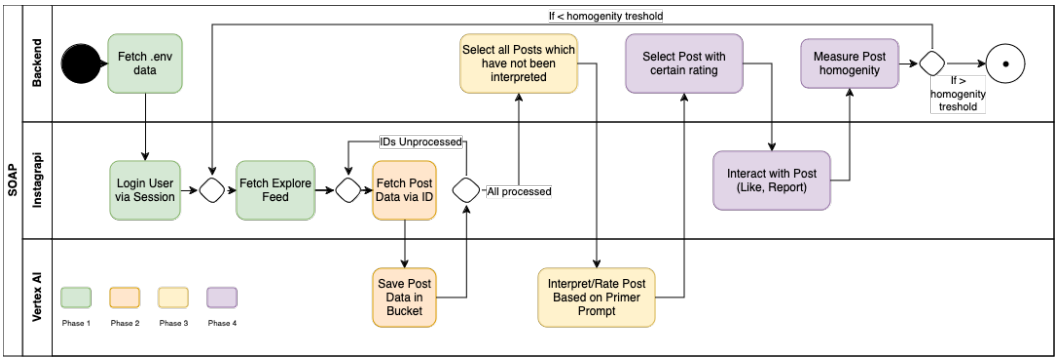


Fig. 2. Workflow of SOAP with its four phases

relevant posts across text, image, and audio modalities. LLMs help triage and structure the data, while human researchers maintain oversight of classification quality (see Section 2.4).

Unlike systems that merely collect content snapshots, SOAP tracks the evolution of personalized feeds over time. This enables causal experimentation, such as comparing the rate at which filter bubbles emerge across different puppet types or measuring how quickly platforms converge toward homogeneity in content. This longitudinal capability provides a distinct advantage over observational API-based datasets. SOAP currently supports Instagram and TikTok and is designed to be extensible to any platform with algorithmic recommendation systems. Its modular design allows researchers to customize puppets, prompts, and engagement strategies to suit diverse auditing objectives. Ongoing development efforts aim to broaden its applicability across additional platforms and regulatory contexts.

3.2 SOAP Workflow

SOAP's operation is divided into four distinct phases (see Figure 2). In Phase 1, the system logs into the VLOP with a user account and opens the user's feed. It fetches the Explore page, which, due to platform defaults, typically returns between 20 and 30 post IDs at a time. In Phase 2, the system processes each of the IDs fetched in the previous step by opening the corresponding posts and collecting associated media files (e.g., videos or images) along with relevant metadata, including the number of likes, username, post text, upload date, and fetch date. In Phase 3, the system utilizes the multimodal large language model (LLM) 'Gemini 1.5 Flash' via the Google Cloud Vertex AI Platform[20] to analyze each post and to identify relevant features or themes in the posts based on the predefined primer prompt. SOAP automatically rates posts according to their relevance with the primer prompt and flags relevant posts. Finally, in Phase 4, based on the analysis outcome, the system employs an interaction mechanism to engage with flagged posts. Similar to user behavior, it may perform actions such as liking, viewing again or saving, depending on how it is configured by the researchers. This process is performed for all fetched posts on the Explore feed page and can be executed continuously for each reload of the feed until the pre-configured homogeneity threshold is reached. This threshold is defined as the ratio of posts matching the filter bubble topic to the total number of posts, indicating when the feed has become predominantly homogeneous.

By adjusting the primer prompt of the multimodal LLM, SOAP can run various sock-puppet audits and therefore explore types of user recommendations and feed constellations. This flexibility

[20] https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/overview Last accessed January 13, 2026.

allows researchers to direct SOAP toward specific topics of interest, as demonstrated in Section 3.3, where the system's validation is discussed.

3.2.1 Data Collection. For the data collection, SOAP uses different publicly available API libraries with OAuth capabilities such as *instagrapi*[21] or *TikAPI*[22]. These are wrappers that interact with social media platforms' private API, enabling both public and private requests, as well as handling challenges during authentication. The libraries allow login by username and password or by session ID. Using these libraries, we can obtain public data of various types, including posts, stories, albums, and Reels. Additionally, this approach provides access to data from the Explore Feed of an account. We retrieve public data such as all comments on a post and lists of users who like a post. Functionalities to interact in a VLOP are also available through these libraries; including liking, following, commenting, saving, and reporting posts via their unique IDs. Notably, the data parameters accessible through these libraries are far more extensive compared to the data provided under the official Research APIs offered by platforms. As discussed in Section 2.3.1, TikAPI, for example, allows the retrieval of over 845 variables, in stark contrast to the 32 variables available through TikTok's research tools. A comprehensive list of all parameters retrieved using different APIs for TikTok, along with descriptions, is available on the GitHub repository of Auditing_TikTok[23].

Since some platforms' Terms of Service (ToS) do not permit automated behavior[24] or the creation of bots/fake accounts, we discuss the legal design considerations in Section 3.4. Platforms actively work to detect, investigate, and block patterns of behavior associated with unauthorized data collection and automated interactions. Hence, to avoid potential account restrictions or bans, we exercised caution in our use of automated tools. The libraries we employed, already include features that reduce the risk of detection by the platforms' automated monitoring systems (e.g., using sessions for logging-in), and we implemented further safeguards (e.g., using location proxies) recommended in the libraries' documentations to minimize activities that might appear suspicious.

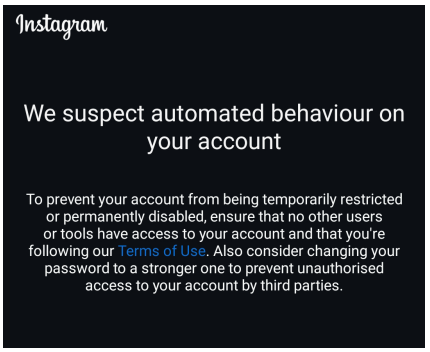


Fig. 3. Notification from Instagram regarding automated behavior on one of the sock-puppet accounts

Despite these precautions, during initial testing, we received warning notifications about detected automated behavior on some accounts (e.g., from Instagram, see Figure 3). We then refined SOAP by incorporating delays of 1 to 3 seconds between interactions to simulate a natural swipe speed. Additionally, we restricted SOAP's activity to a maximum of 300 posts per session, which approximates to two hours of continuous scrolling. In this setup, SOAP replicates the behavior of a

[21] https://github.com/subzeroid/instagrapi. Last accessed January 13, 2026.
[22] https://github.com/tikapi-io/tiktok-api. Last accessed January 13, 2026.
[23] https://github.com/mrtn3000/tiktok-audit/tree/main/Data%20Access. Last accessed January 13, 2026.
[24] See, e.g., https://help.instagram.com/740480200552298. Last accessed January 13, 2026.

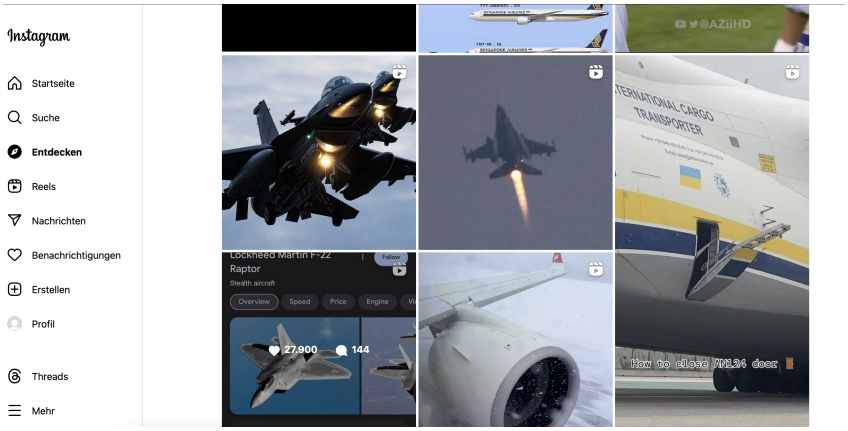


Fig. 4. Highly homogeneous feed of a sock-puppet account trapped in an aviation-related filter bubble (i.e. with aviation-related content homogeneity above a certain threshold).

user engaged in mindless scrolling[25], engaging with content in a continuous, unstructured manner. To create a comprehensive dataset for analysis, SOAP stores all the collected data from an account explore feed, including media content and the associated metadata of each post.

3.2.2 Deductive Coding. For SOAP to interact with posts and the platform beyond random interactions, it must understand and categorize the content it encounters to create highly homogeneous feeds or steer the algorithm in specific directions. To achieve this, the system needs a method to determine whether a post belongs to a particular topic based on a predefined codebook, which includes an initial set of codes, descriptions, and examples aligned with the research focus or theoretical framework [123]. To achieve this, SOAP employs an LLM-based deductive coding approach by including the multimodal LLM *Gemini 1.5 Flash*, which is capable of analyzing video and audio data alongside text. This LLM-based approach significantly enhances scalability and allows for more comprehensive analysis (see Section 2.4).

Acknowledging the limitations of automatic deductive coding (see Section 2.4.1), we remain cautious about relying solely on automated processes and therefore employ cross-validation techniques to ensure the reliability and accuracy of our approach.

3.3 Validation of SOAP

To validate SOAP, we created three sock-puppets on Instagram. Using primer prompts (see Appendix B) and liking as interaction, we steered these sock-puppets toward specific content topics, deliberately inducing filter bubbles characterized by significant reductions in all diversity dimensions as defined by the technological filter bubble [72].

All collected data points and deductive coding interpretations are available on GitHub[26]. We first used SOAP to steer two of the sock-puppets into filter bubbles with little thematic ambiguity, related to Aviation (see Figure 4) and Kittens. Setting a content homogeneity threshold of 75%, the Explore feeds of our sock-puppets reached this level of content diversity after only around 100 (Aviation) or 125 (Kittens) posts (see Figure 5). This corresponds to around 45 to 60 minutes

[25] Mindless scrolling refers to the act of browsing social media or websites for extended periods without a specific goal, enabled by features such as infinite scrolling [92]. For further reference, see https://sites.psu.edu/aspsy/2019/03/16/mindless-scrolling/. Last accessed January 13, 2026.
[26] https://github.com/Interactions-HSG/SOAP

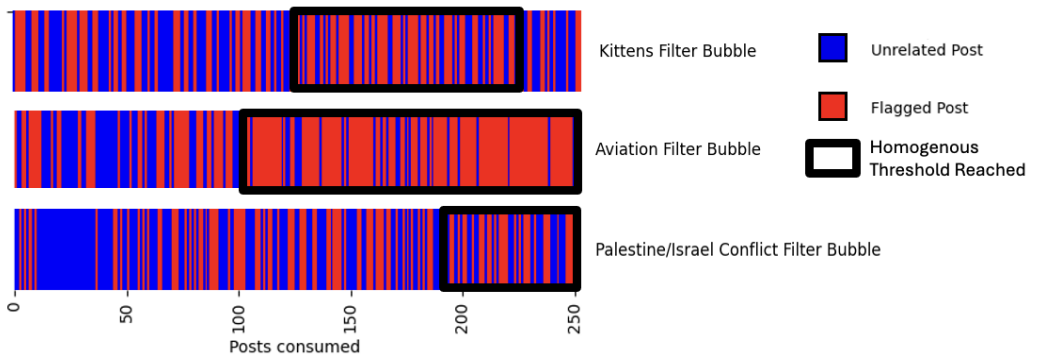


Fig. 5. Content homogeneity of the filter bubbles in which SOAP steered our sock-puppets depicted over time. Here, the homogeneity threshold is set at > 75%.

of scrolling. In the same way, we created a sock-puppet for the more ambiguous topic of the Palestine/Israel conflict. This topic was chosen due to its prominence on social media at the time of writing [99], particularly concerning biased systemic censorship [125], digital activism [18], and war propaganda [2]. As seen in Figure 5, within around 200 posts or roughly 1.5 hours of scrolling, more than 75% of recommended posts were about the conflict, reducing the topic diversity in the Explore feed substantially. To safeguard the integrity of this analysis and address the limitations mentioned in section 2.4.1 , we thus tested the intra- and inter-reliability of the model (i.e., the Gemini 1.5 Flash multimodal LLM) as a labeler.

These reliability checks help to bound the uncertainty of LLM-based coding. Inter- and intra-rater agreement indicate that, for the specified codebook, prompts, and sampled posts, the labeling procedure is stable and internally consistent at the time of data collection. At the same time, agreement scores do not by themselves establish construct validity (i.e., whether the codebook captures the intended concept without systematic blind spots) or external validity (i.e., whether performance generalizes across languages, cultures, time periods, or content genres). We therefore interpret high agreement as support for consistent application of the codebook under our setup, and complement it with targeted spot-checks, topic-specific prompt adjustments, and human validation in sensitive or ambiguous cases.

Intra-reliability Validation. We first assess the consistency of ratings for the same posts through an adjusted Test-Retest Reliability procedure. From each Explore feed of a sock-puppet account, we selected 95 posts and had the model rate each post five times, resulting in a total of 475 ratings per filter bubble. We then calculated Cronbach's Alpha and the 95% confidence interval to quantify the internal consistency of these repeated ratings. The results demonstrate consistently high scores (see Table 1), indicating that the model reliably produced similar ratings across multiple iterations for the same content.

Inter-reliability Validation. We use Human-to-Human and AI-to-Human inter-rater reliability validation to determine if the model's labels align with human labels, a metric previously employed in related studies [19, 20]. Two human labelers and the model independently labeled the same set of 95 posts from each sock-puppet account as used in the intra-reliability validation. Cohen's Kappa [121] was then calculated to compare the ratings between the two human labelers and between the human labelers and the LLM. Table 2 demonstrate the level of agreement, validating the model's reliability in producing labels consistent with human judgment. We observed *substantial*

Table 1. Intra-rater reliability: Cronbach's Alpha and 95% confidence intervals for the model's reliability measured on 95 posts per filter bubble, each rated five time for a total of 475 ratings each.

Filter Bubble	Cronbach's Alpha	95% Confidence Interval
Aviation	0.979	[0.972, 0.985]
Kitten	0.998	[0.997, 0.998]
Palestine/Israel	0.975	[0.966, 0.982]

Table 2. Inter-rater reliability: Cohen's Kappa (κ) for different labelers measured on 95 posts respectively from three different SOAP-generated sock-puppets' filter bubbles. (AI = Gemini 1.5 Flash multimodal LLM)

Labelers	Aviation: κ	Kitten: κ	Palestine/Israel: κ
Human 1 - Human 2	0.9354	0.9794	0.7407
AI-Human 1	0.7705	0.7357	0.4157
AI-Human 2	0.7330	0.7158	0.5042

to *high* agreement in the Aviation and Kitten filter bubbles, and *moderate* to *substantial* agreement in the Palestine/Israel filter bubble. The lower scores for conflict-related content stem largely from topic ambiguity: both human coders and the model sometimes disagreed on whether a post referred specifically to Palestine or Israel conflict. The LLM, in particular, tended to classify generic war-related imagery as "non-related" unless there was an explicit and identifiable link to the Palestine–Israel conflict. In other words, war footage alone was not sufficient for inclusion—the model required contextual cues to assign it to the topic category.

Crucially, even with these lower κ values, the resulting content streams still evolved into clearly homogeneous bubbles (Figure 5). This indicates that near-perfect label agreement is not a precondition for SOAP to generate clear filter-bubble effects. The implications—and limitations—of this observation are further discussed in Section 3.4.4.

3.4 Design Considerations and Limitations of SOAP

The design of SOAP adhered to several considerations stemming from the *legal* framework within the EU, as well as to a range of *ethical* considerations. We discuss these to provide insights into the operation of social media algorithms, enhance transparency, and inform policy decisions.

3.4.1 Legal considerations. In accordance with the regulatory goals of the DSA, SOAP only looks at publicly available data of VLOPs, which is accessible even without an account or user credentials. While [78] suggests that for public interest research there is no 'reasonable expectation' of privacy in social media data that is "publicly accessible in platforms' online interface" , researchers have demanded caution when defining what falls under 'manifestly making publicly accessible (sensitive) data' [27]. With this in mind, we designed SOAP to minimize data collection and analysis, in order to mitigate data privacy risks. Accordingly, SOAP is designed to scrape only publicly available data on Instagram, accessible without an account or login, ensuring compliance with these privacy expectations, as this content circulating in public spaces (i.e., online interfaces of the platform that are available to all users (or potentially also non-users) of the platform), trigger less privacy expectations. Further, SOAP is intended solely for public research purposes and not for any commercial use. As such, these design considerations should not only appease data protection concerns but also about intellectual property ones. In fact, regarding copyright concerns, recent reforms are

easing constraints for researchers. For example, Germany's 2017 provision for text and data mining under § 60d (1) of the German Copyright Act allows for the automatic and systematic collection of data to create a research corpus [77]. Similarly, Article 3 of the EU's Digital Single Market Directive mandates exceptions for text and data mining for scientific research across Member States. These provisions, however, do not grant access to data itself but apply to works to which researchers have lawful access [35, 64]. Lastly, while SOAP was designed for public interest research its goals conflict with VLOPs' ToS (e.g., for the Instagram platform [27]) that prohibit data scraping and automated behavior. In this scope, SOAP interacts with platforms in a natural and unobtrusive way, mimicking standard user behaviors like viewing and liking posts, which are subtle and not readily noticeable to other users. Importantly, SOAP does not engage in active discourse on the platforms—it does not comment, share, or message—ensuring its interactions remain passive. In light of this design, we believe that the impact of SOAP is limited, and the generation of insights on societally relevant subjects, such as the impact of polarization online, should outweigh contractual law considerations possibly brought forward by VLOPs. In fact, from a legal point of view, there is always a need to balance conflicting interests. We see such a balancing of interests in decisions by the European Court of Human Rights deliberating the reach of the "Freedom of Expression" (Article 10 [22]) of the European Convention on Human Rights (ECHR). Two decisions are particularly relevant to situate research like SOAP within the current legal framework: First, in the decision Haldimann and others v. Switzerland[33], the court was asked to balance the breach of a criminal law provision (recording conversations of others without their consent and knowledge) against the need for investigative journalism on a societally relevant issue. Second, in the decision Hertel v. Switzerland [32], the court was asked to balance a potential breach against unfair competition law norms and the publication of a scientific report that is of societal relevance (a report about microwaves' impact on the quality of the food). In both cases, researchers and journalists successfully called upon the "Freedom of Expression" (Article 10 [22]) to defend themselves against an alleged violation of another legal norm. While the first case deals with the collection of data and breaching of the law during the data collection phase, the second case is about how data is reported rather than obtained. We can draw important insights from both such cases, which highlight that a case-by-case analysis is needed when assessing the balance between conflicting legal rights and interests; we argue that research on systemic risks of platforms that is tailored to ensure as much compliance with existing legal norms should remain possible. Notably, the Council of Europe has also recognized the societal importance of such research. In its Declaration on the manipulative capabilities of algorithmic systems, the Council explicitly emphasized "the societal role of academia in producing independent, evidence-based and interdisciplinary research," [76]. This recognition reinforces the importance of safeguarding academic inquiry into the societal impacts of algorithmic systems, particularly where such research serves democratic oversight and public interest.

3.4.2 Ethical considerations. Regarding ethical considerations, SOAP has the capability to generate feeds containing extreme content (e.g., violence). Although platforms such as Instagram assert that they remove such material, numerous audits—as well as our own research—demonstrate that disturbing content continues to circulate on the platform [53, 62], including content that should have been removed under the platforms' own content moderation policies. This persistence raises concerns regarding the protection of content moderators and researchers who may be exposed to harmful material when using SOAP or during the application of deductive coding schemes. Such exposure poses significant risks to the mental health and well-being of those involved [106]. To mitigate these risks, SOAP employs *automatic* deductive coding using multimodal LLMs, thereby reducing the need for direct human exposure to potentially harmful content (cf. [20]).

[27] https://help.instagram.com/740480200552298. Last accessed January 13, 2026.

Another, related, ethical consideration is that we cannot guarantee that actors will not use SOAP for unintended purposes. As a safety and precautionary mechanism, SOAP includes safeguards to prevent harmful and not-safe-for-work (NSFW) prompting (see Section 2.4.1). However, especially due to the open-source nature of our framework, these safeguards may be circumvented by replacing the deductive coding component. In this context, SOAP exists within the same ethical field of tension as many other AI-enabled tools [103].

3.4.3 Platform Restrictions. As mentioned in Section 3.2.1, VLOPs implement numerous restrictions and mechanisms to prevent automated behavior and data scraping, often for reasons beyond research concerns. However, by using API libraries like Instagrapi and TikAPI in a deliberate and cautious manner, some of these restrictions can be circumvented. Even if these libraries provide methods to bypass certain barriers (see Section 3.2.1), they remain vulnerable to platform countermeasures, such as API configuration changes or increased scrutiny of automated accounts. For SOAP, this means that some sock-puppet accounts, despite careful design and adherence to ethical and legal standards, may be flagged, restricted, or even banned during operation. While such challenges are understandable given the potential misuse of these tools for malicious activities—such as creating bot armies[28], or executing Collusion Attacks, Ballot-Stuffing, Whitewashing, Sybil Attacks, and other forms of automated manipulation [30]—their careful application in research remains crucial.

Instagrapi and similar libraries face constant risks from the technical and legal barriers imposed by platforms. Ultimately, a legal framework should support the use of such tools, especially in legitimate use-case scenarios like academic research.

3.4.4 Limitations of Automated Deductive Coding. SOAP leverages sock-puppet accounts and multimodal LLMs to explore filter bubbles by emulating user interactions with personalized recommender systems. While this approach provides a scalable way to investigate how VLOPs respond to specific interest profiles, it is essential to clarify its epistemic and methodological boundaries.

Firstly, SOAP does not aim to replicate genuine human behavior. Sock-puppet accounts follow predefined, structured interaction patterns designed for consistency and replicability—not for behavioral realism. They do not reflect the full spectrum of user behaviors, nor do they account for the complex personal, social, and contextual signals typically used in personalization algorithms, such as social graphs, emotional responses, or multi-platform activity. As a result, while SOAP reveals how recommender systems react to controlled inputs, it does not simulate the full complexity of real-world user engagement. Filter bubble-like effects observed in SOAP should therefore be understood as system-level responses to stylized probes, not as representative of what actual users might experience.

Secondly, the deductive coding process within SOAP has topic-specific limitations. As discussed in Section 2.4, we use a multimodal LLM to classify whether posts correspond to predefined thematic categories (e.g., aviation, kittens, or the Israel–Palestine conflict). This approach enables large-scale labeling across multimodal content, but it faces challenges in ambiguous or sensitive domains. In particular, we observed lower inter-rater agreement for conflict-related content—partly because the model only labeled a post as relevant when it contained clear references (e.g., to Gaza or Israel), excluding more generic war imagery. These edge cases led to lower Cohen's κ scores (see Section 3.3). Nevertheless, the resulting personalized feeds still displayed strong thematic convergence (Figure 5), suggesting that SOAP can create filter bubbles even when coding precision

[28] See, e.g., https://www.forbes.com/sites/siladityaray/2024/06/05/israel-reportedly-used-fake-social-media-accounts-to-garner-support-from-us-lawmakers-on-gaza-war/ or https://www.wired.com/story/russian-influence-campaign-exploiting-college-campus-protests/. Last accessed January 13, 2026.

is imperfect. Still, such findings should be interpreted with caution and supported by human validation in more ambiguous topics.

Thirdly, LLM integration in SOAP is deliberately narrow and empirically validated. We use the LLM strictly for topical classification, not for simulating user preferences, decisions, or affective responses. Its outputs are tested through intra & inter-rater validation to ensure consistency with human annotations, and its use is fully documented through shared prompt templates and evaluation procedures. Given known limitations of LLMs—including biases in politically charged contexts, reduced accuracy across languages, and the inability to process recent or underrepresented topics—we treat the model's output as a structured approximation, not a substitute for human interpretation. We encourage future researchers using SOAP to conduct complementary human annotation, adapt prompts to specific domains, and remain attentive to the known boundaries of LLM performance. Overall, SOAP uses multimodal LLM coding to scale *descriptive* labeling for comparative audits. It is not a substitute for human judgment. We therefore interpret coded outputs as indicators of exposure dynamics under controlled probes and recommend targeted human validation whenever conclusions depend on fine-grained distinctions or sensitive content interpretations.

In summary, SOAP is not intended to replace user studies, data donation efforts, or crowd-based annotations. Rather, it provides a structured, reproducible, and legally compliant method for investigating systemic risks in VLOP recommender systems. Its findings must always be contextualized within the limitations of its simulated behavior and automated coding pipeline.

4 Applications of SOAP

With its ability to analyze recommendation dynamics and simulate diverse user inputs, SOAP allows researchers to investigate indicators and pathways related to systemic risks under controlled conditions. It addresses questions that remain challenging to tackle through existing data access methods under the DSA or conventional research APIs, offering a framework for studying how personalized recommendation mechanisms can shape content exposure in ways that are consistent with systemic-risk concerns.

SOAP is designed to support comparative and mechanistic analyses of personalization: for example, how content exposure differs across personas, how recommendation pathways evolve under controlled interaction policies, and how platform settings affect the recommendations shown to specific account types. At the same time, a controlled sock-puppet setup cannot on its own establish population-level prevalence, attribute observed outcomes causally to platform-wide policies without additional experimental designs and controls, or directly measure real-user experiences and downstream harms. Accordingly, findings derived from SOAP should be interpreted as observations from controlled accounts that can flag potential risks and motivate follow-up studies (e.g., triangulation with data donation, crowd-based audits, or human-in-the-loop evaluations).

Beyond research, SOAP has been successfully employed in workshops to raise awareness about filter bubbles and the societal impacts of personalized recommender systems. These workshops bridge theoretical insights with practical applications, equipping participants to critically engage with algorithmic systems and understand their influence on public discourse and societal polarization.

4.1 Auditing Systemic Risks Using SOAP

Building on existing methodologies for auditing systemic risks under the DSA (see Section 2.2), SOAP provides a framework for addressing key limitations in current approaches to measure such risks. In the following, we outline initial applications and propose directions for how SOAP can be used to audit systemic risks more effectively. While these examples illustrate the tool's potential,

we acknowledge that some remain conceptual at this stage. Our study demonstrates the technical feasibility of the approach and outlines research pathways that need further empirical investigation.

4.1.1 Dissemination of Illegal Content. SOAP enables comprehensive analysis of how illegal content, such as posts inciting violence, promoting hate speech, or containing libel is disseminated on VLOPs. By interacting with this content, SOAP captures the pathways through which recommendation algorithms may promote or amplify these materials, offering a detailed view of their life-cycle from creation to amplification.

Recent investigations, such as the Wall Street Journal's report on Instagram's facilitation of a vast pedophile network, underscore the urgent need for tools like SOAP to scrutinize platform mechanisms. The report revealed how Instagram's algorithms and community-building systems inadvertently connected users with harmful interests, raising significant concerns about the role of recommendation systems in fostering and amplifying illegal activities[29]. Additionally, The Guardian reported that Instagram failed to remove accounts featuring concerning images of children, even after they were reported by users[30]. Moreover, research has shown that illegal and harmful content often resides at the "tail of the distribution," where it is actively sought out by a narrow, motivated group of users [15]. While public discourse often emphasizes high exposure to false or inflammatory content, empirical evidence suggests that average exposure is low and concentrated among those with strong motivations to seek such information. Filter bubbles, in this context, can be self-imposed, with users deliberately curating their feeds to reinforce specific ideologies or political interests [29], including exposure to illegal content. This phenomenon raises critical questions about the role of algorithms in facilitating such exposure. Although personalized recommender systems may not be the primary cause of harmful content consumption, they can exacerbate the problem by tailoring recommendations that further entrench users within these harmful niches [15]. SOAP's ability to automatically interact with content allows for a nuanced investigation of how illegal content is distributed and amplified within algorithmically reinforced filter bubbles. Additionally, SOAP enables the measurement of exposure and mobilization among extremist fringes, focusing on tail exposure metrics that capture interactions with false and extremist content.

Research APIs, in contrast, often restrict access to flagged or sensitive content, significantly limiting researchers' ability to study these issues. Such restrictions prevent the examination of the full lifecycle of harmful posts, from their initial appearance, to their recommendation in the user feed and their (eventual) removal or suppression. SOAP addresses this gap by logging detailed content interactions to capture the algorithmic mechanics for content dissemination.

4.1.2 Negative Effects on the Exercise of Fundamental Rights. SOAP allows for the examination of biases in content moderation practices and their broader implications for representation and non-discrimination; by mimicking interactions that represent a variety of demographic perspectives, SOAP could be used to uncover disparities in how content is moderated or recommended, providing insights into potential systemic biases.

SOAP also facilitates the detection and analysis of shadowbanning by logging engagement metrics and identifying patterns where certain content or demographics experience unexplained suppression. As platforms often deny the existence of shadowbanning when users report its effects [26], SOAP offers an evidence-based approach to uncover discrepancies. For instance, platforms lack transparency in communicating moderation decisions or metrics, making users

[29] https://www.wsj.com/tech/instagram-vast-pedophile-network-4ab7189. Last accessed January 13, 2026.
[30] https://www.theguardian.com/society/2022/apr/17/instagram-under-fire-over-sexualised-child-images. Last accessed January 13, 2026.

develop "folk theories" on algorithmic suppression [26]. SOAP can systematically investigate these occurrences, providing empirical data to validate or challenge user experiences and platform claims.

4.1.3 Negative Effects on Civic Discourse, Electoral Processes, and Public Security. SOAP is particularly well-suited to studying algorithmic influences on civic discourse and democratic processes. By interacting with political content through sock-puppet auditing, SOAP provides a unique capability to analyze how recommendation systems serve political content across different personas and ideological perspectives. For instance, it enables researchers to measure the extent to which political content is prioritized, suppressed, or amplified for users with varying political leanings. This insight into differential content exposure is critical for understanding how algorithms shape public opinion and influence political engagement. Instagram, for example, has introduced a policy to throttle political content recommendations in users' Explore Feeds, as part of an effort to limit the spread of political material. Under this policy, political content is reduced by default and must be enabled manually by users if desired [83]. SOAP enables researchers to empirically assess such platform-imposed content restrictions, providing insights into how they influence the visibility of political content and affect user experience.

Additionally, SOAP supports studies that measure the diversity of content exposure, addressing the concerns highlighted by Calabrese et al. [17]. This includes identifying how recommendation algorithms may create ideological filter bubbles or echo chambers that distort public discourse. By performing real-time studies during politically sensitive periods, such as elections, SOAP can observe how recommendation systems adapt and potentially amplify polarizing narratives under societal tension. This approach has already been applied in the context of the 2025 German federal election, where SOAP was used to investigate the content recommended to users expressing interest in the AfD party on TikTok [74]. These capabilities are essential for analyzing how systemic risks emerge and persist during critical moments of democratic stability.

4.1.4 Negative Effects Related to Gender-Based Violence, Public Health, the Protection of Minors, and Individuals' Physical and Mental Well-Being. SOAP facilitates research into how algorithmic recommendations impact vulnerable populations. Recent reports have highlighted alarming systemic risks stemming from personalized recommendation systems, particularly concerning vulnerable populations such as minors and individuals at risk of physical or mental harm. Amnesty International, e.g., demonstrated how TikTok's "For You" feed promotes self-harm and suicidal ideation [53]. Similarly, projects *Recommending Toxicity* [5] and *Safer Scrolling* [87] revealed how YouTube Shorts and TikTok amplify male supremacist influencers and online hate, gamifying misogynistic content for young users. An EU report [86] further underscores the troubling influence of social media platforms on mental well-being and societal perceptions, particularly for women and girls. In response, platforms like Instagram claim to mitigate these risks through measures such as Teen Safety settings[31], which are purportedly designed to block harmful content for underage users. Additionally, the Digital Services Act (DSA) in the European Union mandates protections for minors under Article 28(2), prohibiting personalized advertising for users under 18 years of age. These platform-provided safety features are frequently highlighted in transparency reports as evidence of their commitment to user protection. However, these claims remain largely unchecked due to limited external scrutiny and restricted access to platform operations.

Currently, there is no robust mechanism for independently validating these claims or evaluating the efficacy of these protections. Platforms often control the narrative by selectively disclosing information, leaving significant gaps in understanding the real-world impact of their algorithms on vulnerable groups. SOAP addresses this critical gap by enabling researchers to conduct controlled

[31] https://about.instagram.com/blog/announcements/instagram-teen-accounts. Last accessed January 13, 2026.

investigations through sock-puppet audits with simulated child or teen accounts. These audits enable researchers to systematically study the content and advertisements shown to minors on these platforms. For example, SOAP can be used to assess whether simulated underage accounts are exposed to targeted advertisements which would directly contravene EU regulations prohibiting personalized ads for minors. SOAP's framework has also been used by Swiss public broadcaster SRF in a journalistic investigation of TikTok recommendations in the online "manosphere" [3].

4.1.5 Overall Policy Insights. Providing adequate oversight over systemic risks on VLOPs requires robust investigative tools. SOAP offers one such instrument, designed to support regulators, researchers, and civil society actors in auditing how recommender systems shape content exposure. While we acknowledge in Section 3.4.4 that SOAP does not replicate the full behavioral and cognitive complexity of real users, the tool nonetheless provides valuable system-level insights. Specifically, SOAP enables structured, repeatable audits of platform behavior under controlled conditions, offering regulators empirical evidence of how recommendation engines respond to clearly defined interest profiles.

Even when behavioral realism is limited—e.g., when simulating children's accounts or multilingual users—the detection of certain content patterns can carry policy relevance. For instance, if a simulated child user is exposed to personalized advertising or harmful content types (as outlined in Section 4.1), this may indicate non-compliance with legal safeguards under the DSA and related child protection regulations. In such cases, the focus is not on perfectly modeling a child's browsing behavior, but on observing whether prohibited outcomes emerge under testable conditions.

We therefore view SOAP as a complementary tool that regulators can use to flag potential systemic risks. Its findings can motivate deeper investigations, inform risk assessments, and help evaluate the credibility of VLOP transparency reports. However, we stress that insights from SOAP should be interpreted in context—as indications of platform affordances rather than as representative of all user experiences. As such, we encourage a mixed-methods approach that includes data donation, crowd-based studies, and human-in-the-loop auditing to triangulate systemic risk analyses.

4.2 SOAP as an Awareness Tool in Workshops

One strategy for tackling systemic risks is to increase media literacy, helping users recognize and avoid filter bubbles. This involves educating users on the importance of diversifying their media sources and verifying information through multiple channels [17]. Research emphasizes the importance of general education, digital skills, and critical media literacy as crucial factors in mitigating vulnerabilities to disinformation and online manipulation, which are key systemic risks tied to personalized recommender systems [9, 75]. Importantly, there is a growing need to incorporate AI literacy, as machine learning and algorithmic content curation become increasingly central to how users engage with information online and critical media literacy is quite low in certain parts of societies across Europe [38]. Incomplete understanding of these technologies exacerbates vulnerability to persuasive techniques like Political Microtargeting (PMT), potentially distorting civic discourse and electoral outcomes [61]. For example, understanding of subjective persuasion techniques was found to predict skepticism toward PMT in a sample of voters from the United States [71]. By strengthening the development of digital skills and critical media literacy—particularly in relation to filter bubbles and personalized recommender systems—individuals can more effectively navigate complex online environments, critically evaluate the information they encounter, and reduce their susceptibility to disinformation and manipulative content. AI literacy extends this foundation by addressing the ethical challenges posed by algorithmic systems, particularly those influencing mass personalization. As Hermann [50] highlights, explicability, in the form of intelligibility, is a prerequisite for individuals to critically assess biases, agency,

privacy concerns, and benefits associated with AI-driven personalization. A lack of understanding leads to a reliance on opaque systems, which undermines judgments about justice, autonomy, and non-maleficence. AI literacy could empower individuals to judge beneficence, non-maleficence, autonomy, and justice related to AI-driven mass personalization themselves. AI literacy could further help individuals to retain autonomy and agency, since they are better able to identify and assess choice architectures generated through AI-driven mass personalization [50]. Such literacy empowers individuals to engage critically with algorithmic content, thereby mitigating systemic risks like polarization and the erosion of trust in civic processes.

These insights underline the critical importance of enhancing media, digital, and AI literacy to empower users in navigating the complexities of personalized recommender systems. By fostering awareness and critical engagement, individuals can better identify and mitigate systemic risks such as filter bubbles, misinformation, and polarization. However, addressing these challenges requires not only theoretical understanding but also practical tools and frameworks that can effectively educate and inform users. SOAP, with its ability to create controlled filter bubble environments, offers a tangible method for raising awareness and fostering critical literacy in this domain. Building on the identified need for enhanced media, digital, and AI literacy, the following report of a workshop demonstrates how SOAP serves as an effective tool for education and awareness. Designed to engage users directly with the dynamics of algorithmic recommendations and filter bubbles, the workshop highlights SOAP's potential to bridge theoretical insights with practical learning. By immersing participants in conspicuous filter bubbles, the study evaluates their awareness of systemic risks and their ability to critically reflect on the influence of social media algorithms and their algorithmic media content awareness.

4.2.1 Method. During the kick-off days for incoming Master and PhD students at the University of St.Gallen in September 2024, we conducted a workshop titled "Behind the Feed: The Influence of Social Media Algorithms on Public Opinion" as part of the academic program. It was one of nine workshops the students could choose. The kick-off days' academic program was framed by the topic "Big, Great Challenges" and brought together interdisciplinary perspectives to demonstrate how these challenges can be tackled.[32] Our workshop aimed to educate the participants about social media algorithms and filter bubbles, especially in the context of political micro targeting in the US election. Additional to the educational aspect of the workshop, we used this opportunity to record the participants opinions and experiences with social media. This allowed us to study user behavior and perceptions after exposure to clearly visible filter bubbles created by SOAP — contrasting with the subtler filter bubbles users may experience in their own social media use. The study thus aimed to assess whether exposure to these conspicuous filter bubbles influences users' awareness of algorithmically recommended content and the presence of filter bubbles on social media platforms.

In the workshop, we first introduced participants to the current challenges in social media, emphasizing the conflicting values identified by Stray et al. [110]—such as balancing free expression and safety, safeguarding privacy while ensuring usefulness, addressing short-term needs versus fostering long-term benefits, and navigating between harmful and constructive forms of connecting. Afterwards, we asked the participants to fill out a pre-survey which asked participants about their perception on current social media. Additionally, the Algorithmic Media Content Awareness (AMCA) scale [127] was used to assess the extent to which the participants understand the role of algorithms in selecting and presenting media content. We adjusted the items slightly to fit them in our context (see Appendix D.1). The scale captures four key dimensions:

[32] See https://www.unisg.ch/en/studying/starting-your-studies/hsg-kick-off-days/academic-programme/. Last accessed January 13, 2026.

(a) (b) (c)

Fig. 6. (a) A screenshot of a Democrat-leaning filter bubble generated by SOAP as used in the workshop , and photos from the workshop showing (b) the participants interacting with an such an Instagram account created by SOAP, and (c) the participants of one group during the group challenge.

- *Content Filtering Awareness (FIL)*: Users' awareness that algorithms filter and personalize content based on their online behaviors.
- *Automated Decision-Making Awareness (ADM)*: Users' awareness that algorithms are making automated decisions about the content they see, without direct human involvement.
- *Human-Algorithm Interplay Awareness (HAI)*: Awareness that users' own behaviors and data influence the algorithms and, consequently, the content presented to them.
- *Ethical Considerations Awareness (ETH)*: Users' awareness of the ethical implications of algorithmic content, such as privacy issues, bias, and transparency concerns.

Following an introduction to filter bubbles on social media and the SOAP tool, participants engaged with Instagram accounts created using SOAP. These accounts, or bots, were designed to simulate filter bubbles and were presented through the standard Instagram user interface on large screens distributed throughout the workshop room (see Figure 6). Each group of 10–15 participants interacted with the bots for approximately 20 minutes, exploring feeds that were intentionally curated to represent U.S. election filter bubbles.

From July to September 2024, the bots were run on Instagram to explore specific filter bubbles. Over several months, ten bots interacted with and collected data from their Instagram Explore feeds. These bots operated based on a U.S.-election-focused primer prompt, deliberately designed to remain neutral to assess variations in viewpoint, topic, and structural diversity intensity. The primer prompt (see Appendix B.4) ensured a controlled starting point for evaluating how Instagram's algorithms would develop the filter bubbles.

To facilitate the creation of distinct filter bubbles and explore different scenarios, the bots were configured with varying behaviors. Some followed politically aligned accounts from across the spectrum, while others followed no accounts at all. This approach allowed us to observe how differing levels of interaction influenced the evolution of content diversity within their feeds. The political spectrum of the bots, based on the accounts they followed, is illustrated in Figure 7. A comprehensive list of the usernames and followed accounts of the bots is provided in Table 4 in Appendix C. After the interaction with the Instagram accounts, participants completed a post-survey designed to capture their experiences with social media, filter bubbles and their perceptions of the content they had just seen. Also, the AMCA scale was included again, so that we could later assess changes in the participant's answers between the pre- and post-survey.[33] After the second survey, the participants stayed in the same groups as before and were tasked with the challenge to find a solution to the question "How can we address the risks and challenges associated with

[33]The surveys are provided in Appendix D.1 and D.2.

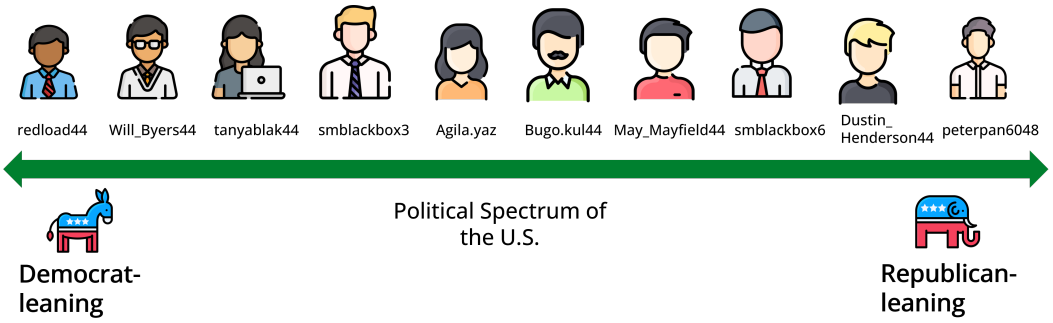


Fig. 7. Usernames of the bot accounts used in the workshop across the political spectrum of the U.S., sorted from Democrat- (left) to Republican-leaning (right) based on the followed accounts (see Appendix C). Person icons sourced from Flaticon (https://www.flaticon.com/authors/vitaly-gorbachev).

Table 3. AMCA statistics (5-point Likert scale) and Wilcoxon signed-rank test results (N=100). All mean differences per dimension and combined are significant. (1 = "Not at all aware", 5 = "Completely aware")

Dimension	Pre-Survey: Mean (SD)	Post-Survey: Mean (SD)	W	p-value	Rank-Biserial Correlation r
FIL	4.59 (0.50)	4.77 (0.40)	270.5	<0.001	-0.538
ADM	4.17 (0.70)	4.53 (0.61)	149	<0.001	-0.742
HAI	4.37 (0.65)	4.58 (0.63)	277	<0.001	-0.400
ETH	4.05 (0.71)	4.46 (0.65)	341.5	<0.001	-0.649
Combined	4.32 (0.49)	4.60 (0.46)	321.5	<0.001	-0.750

the influence of social media algorithms on public opinion?". Afterwards each group presented their solution to all workshop participants. The workshop and data collection were exempt from a formal review by the Ethics Committee of the University of St.Gallen.

Participants. Our workshop was attended by 155 participants. 144 of them filled the pre-survey and consented to the data usage, while 120 did the same for the post-survey. After excluding participants who completed none or only one of the surveys, or did not consent to data usage, the answers of 100 participants were evaluated. Given the size of the workshop, we did not enforce the completion of the surveys. The evaluated participants were on average 23.66 years old (SD=2.24), and 44 participants identified as women, and 56 as men. Most were beginning their Master's programs in Accounting and Corporate Finance (N=36), Banking and Finance (N=11), and Marketing Management (N=11), followed by General Management (N=9), International Affairs and Governance (N=7), and Strategy and International Management (N=4). Four participants joined the workshop that were about to start their PhD. Most participants indicated using the social media platforms Instagram (N=92) and LinkedIn (N=92), followed by Snapchat (N=63), TikTok (N=40), Facebook (N=34), and Twitter/X (N=17; see Table 6 in Appendix E for the full list of details). Furthermore, 71 participants had seen political content recently in their social media feeds, while 21 did not, and eight were unsure.

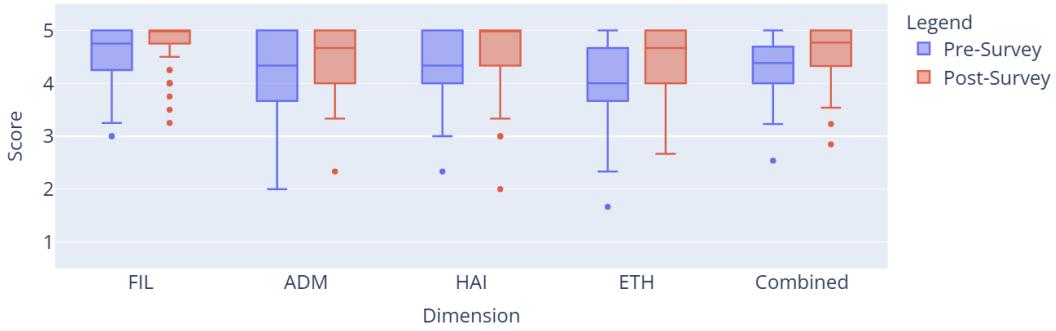


Fig. 8. Boxplots of the AMCA results in the pre- and post-survey per dimension and combined (N=100). (1 = "Not at all aware", 5 = "Completely aware")

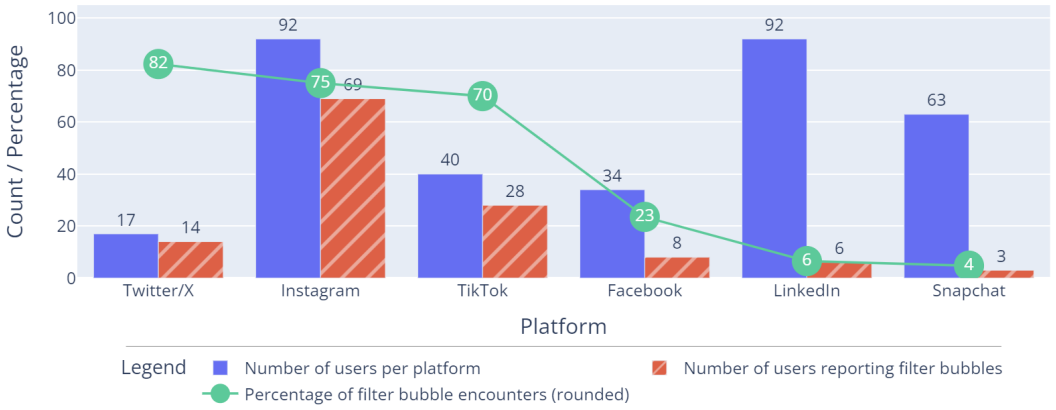


Fig. 9. Filter bubbles in social media. The chart above shows the following: (1) total number of users per platform (blue bars). (2) number of users reporting filter bubbles (orange bars): Participants self-reported whether they had experienced a filter bubble on each platform. (3) percentage of filter bubble encounters (green line): The percentage of users per platform who reported entering a filter bubble, with Instagram, X/Twitter, and TikTok leading, as reflected by the green line

4.2.2 *Results.*

AMCA. In order to compare the AMCA scale results of the pre- and post-survey, we conduted Wilcoxon signed-rank tests because the data were paired and did not meet the assumption of normality for a parametric test. The results indicated a significant improvement from the pre- to the post-study for the combined dimensions ($W = 321.5, p < 0.001, r = -0.75$), as well as for the individual dimensions (see Table 3 and Figure 8).

Filter Bubble Experience. Most participants encountered filter bubbles in their own social media feed recently (N=79). A majority of them reported encountering filter bubbles "Very frequently" (N=60), some "Occasionally" (N=17), and only two "Rarely". The platforms with the highest percentage of self-reported filter bubbles were Twitter/X, Instagram, and TikTok (see Figure 9). Interestingly, despite being heavily used, platforms like LinkedIn and Snapchat showed lower occurrences of filter bubbles as reported by users.

Indicators of Filter Bubbles. In the pre-survey the participants were asked for indicators that a feed is highly curated or biased, and in the post-survey whether they noticed any indicators that suggested the content of the account they analyzed was part of a filter bubble. We categorized their answers into the diversity dimensions of filter bubbles proposed by Michiels et al. [72] (one answer can be in multiple dimensions).

In the pre-survey, almost half of the participant's comments were connected to viewpoint diversity (N=40), such as a lack of different viewpoints in the presented content, or content that is predominantly based on ones' own viewpoint, e.g., P96 perceived "almost no different viewpoints displayed (e.g. on X)" as an indicator, and P59 remarked that "[since] social media only shows you idea you agree with, it shows how biased algorithms are, which can be dangerous". Around half of the participants in the post-survey then noticed viewpoint related indicators when analyzing the SOAP-created feeds (N=48), e.g., P21 observed that "[the] posts were all in favor of one party and they were really similar and lacked nuance", and P56 saw "[only] posts in link with the same political opinion as the person. Nothing contradictory".

Around a third of the participants connected to the topic diversity in the pre-survey (N=30), remarking on the same content being presented repeatedly and on content that is presented based on their interactions on the platform, e.g., P43 saw indicators for highly curated feeds when "it is similar to other posts i've interacted with before", and P18 when "[it] displays always the same kind of content". Again, more participants commented on topic-related indicators after being exposed to the SOAP-generated feed (N=42), e.g., P63 perceived that "everything was about politics" and P29 saw the "same topics and people throughout the feed".

Only a few participants remarked on the structural diversity (pre: N=15, post: N=9), such as a lack of sources, or much content coming from one or few identical sources, e.g., P58 (pre) saw "Unreliable sources or no sources at all" as a indicator for a biased feed, and P28 (post) found that the "comment sections which I normally consider as a 'safety net', was not showing different views and was fully in line with the view or content of the reel/post" when comparing the SOAP-generated feed to their usual social media feed.

In the pre-survey, the participants commented also on other indicators for highly curated and biased feeds, such as personalized advertising based on previous interactions on the platform and other Websites, highly edited photos and videos, or clickbait titles.

In the post-survey the participants also commented on their general perception of the feed in connection with their knowledge on algorithms, e.g., P34 perceived the content in the feed as "totally a bubble which is pretty shocking even if I already were aware of the [algorithms]", and P22 could "observe a [reiterating] algorithm" in the feed's content.

Comparison of SOAP-Created Filter Bubbles with Participant Experiences. The participants rated the tone and content of the SOAP-feeds significantly and strongly more partisan than their own social media feeds ($W = 191, p < 0.001, r = -0.972$; see Table 7 in Appendix E). Likewise, they rated the likelihood of being influenced by the SOAP-feeds if they were consistently exposed to it, significantly higher than for their own feeds ($W = 782, p = 0.003, r = -0.854$). Also, participants found the information in the SOAP-feeds significantly less credible than in their own feed ($W = 642, p < 0.001, r = -0.882$). All rank-biserial correlations r indicate strong effects.

In a free-text field in the post-survey, the participants were asked to compare the SOAP-feeds to their typical social media feed. Given that SOAP was designed to create heavily homogeneous and extreme filter bubbles, in this case particularly around the topic of the U.S. election, we anticipated that participants' personal feeds would differ from the SOAP-generated content. We clustered the answers in same three diversity clusters as above, and summarized recurring themes:

Viewpoint Diversity Differences. One-third of participants observed that the SOAP-generated feeds were more biased, extreme, or one-sided than their personal feeds (N=28). They noted that the content was heavily opinionated or polarized, which differed from the more balanced or neutral content they usually encounter. P81, e.g., observed that "the filter bubble [i.e. SOAP feed] had extremer, opinion-oriented content" and P44 perceived the analyze feed as"[...] very one sided. I think mine [is] more neutral ".

Several participants observed that the political leanings in the SOAP feeds did not align with their own (N=7). They noted differences in the representation of political figures and the overall political orientation of the content, e.g., P63 commented that the SOAP feed "[...] was right wing oriented, mine is more liberal", and P12 admits "[my] feed only portrays trump in a bad light and is very pro Kamala".

Interestingly, a group of participants reported that the SOAP-created filter bubbles were similar to some extent to their personal social media experiences (N=9). Despite the expectation that the extreme and homogeneous nature of the SOAP feeds would differ from most users' experiences, these participants indicated that they regularly encounter comparable content, e.g. P98 describes the SOAP feed as "almost the same as mine", and P27 states: "It's pretty much the same, a strong support for Trump and a [bad representation] of Biden and Harris". In contrast, almost twice as many participants stated that the SOAP feeds are completely different to their own feeds (N=14).

Topic Diversity Differences. Around one third of the participants noted that the SOAP-created filter bubbles contained more political content than their own feeds (N=29). Many participants mentioned that they typically encounter less political content, focusing instead on other topics such as lifestyle, sports, or entertainment. P1, for instance, describes their experience: "I'm not really into politics so I basically never get these political posts but I do get posts always in the same category such as pets, relationship, fashion but I wouldn't say that it's a filter bubble but rather a topic filter", and P70 explains: "In my social media feeds I don't have political information but more about sports, nutrition, influencers, lifestyle, etc". Approximately one fourth of the participants mentioned that the SOAP feeds were less diverse, often focusing intensely on a single topic or viewpoint (N=22). This contrasted with their own feeds, which they described as more varied and covering a broader range of interests, e.g., P7 states: "I feel like my feed is much more diverse, with very various contents.", P8 describes their feed as "[...] more balanced and fact based", and similarly P48 sees their feed as "absolutely mixed & diverse".

Structural Diversity Differences. Only few participants (N=5) commented on the structural diversity (i.e. the variety of information sources), such as P52 who states: "my bubble is more politically biased, [I] only follow one party", or P82: "I follow a lot of news agencies like the NYT, WSJ, FT, and Washington Post, which are not neutral messages, but are more credential".

Group Challenge. Notably, four out of ten groups proposed partially replacing personalized recommendations with randomized or unbiased content in users' feeds. For instance, some groups suggested that 20-25% of the feed should include either randomly generated content or content curated to reduce polarization. Moreover, participants perceived platforms as bearing the primary responsibility for implementing such changes, with eight out of ten groups emphasizing the need for platform-level algorithmic adjustments. For instance, suggestions included creating dynamic algorithms that actively steer users toward diverse viewpoints or broadening recommendation inputs by incorporating the activities of a user's social connections. Participants also acknowledged the importance of user education, with two groups proposing initiatives to improve media literacy and raise awareness of algorithmic influence. Proposed measures included in-app educational tools and programs targeting younger audiences to teach them about the risks of filter bubbles and the

importance of diverse perspectives. Furthermore, some groups suggested enhancing transparency through features like political algorithm settings or pop-up warnings before displaying potentially biased or polarizing content.

4.2.3 Discussion of the Workshop Results. In this workshop, we created filter bubbles using SOAP's sock-puppet accounts, simulating user behaviors to produce highly curated feeds centered on the US election. These filter bubbles allowed participants to directly experience and reflect on the potential systemic risks of personalized recommender systems, such as the spread of misinformation and political polarization. Through our evaluation we gathered insights into participants' perceptions of these curated feeds and their understanding of algorithmic influence.

AMCA. The significant results for the AMCA scales indicate a marked improvement in participants' algorithmic content awareness after interacting with SOAP-generated filter bubble user feeds. The increase in awareness was observed across all dimensions of the AMCA scale. This suggests that interaction with SOAP effectively enhances participants' understanding of the mechanisms and implications of algorithmic filtering and personalized recommender systems. As discussed in Section 4.2, heightened algorithmic literacy—combined with awareness of subjective persuasion knowledge [71]—can foster greater skepticism toward practices like political microtargeting. Consequently, this increased literacy could mitigate the impact of filter bubbles on public opinion, contributing to more informed and critical engagement with algorithmically curated content.

Filter Bubble Experience. The findings highlight the significant presence of filter bubbles on platforms such as Instagram, X/Twitter, and TikTok, where a majority of participants reported encountering filter bubbles in their usage. This underscores the strong influence of personalized recommender systems in shaping user experiences and limiting exposure to diverse perspectives. Interestingly, platforms like LinkedIn, and Snapchat showed comparatively fewer reported filter bubble encounters. This may indicate differences in their algorithmic design, levels of personalization, or the nature of content curation. For example, LinkedIn's focus on professional and academic networks may inherently create a bubble based on users' educational or career environments, which participants may perceive as neutral or not immediately recognize as a filter bubble. Similarly, Snapchat's emphasis on ephemeral, direct communication [88] might limit the persistent curation of content that fuels filter bubble formation. In contrast, platforms like Instagram, TikTok, and Twitter/X, which prioritize algorithmically curated, high-engagement content, appear more prone to fostering noticeable filter bubbles. Overall, the large number of participants reporting filter bubble experiences confirms the widespread nature of this phenomenon, particularly on platforms where algorithmic personalization dominates content delivery.

Comparison of SOAP-Created Filter Bubbles with Participant Experiences. As anticipated, the analysis revealed that the SOAP-created filter bubbles exposed participants to content that was notably more biased, political, and homogeneous compared to their personal social media feeds. Many participants highlighted the extreme and one-sided nature of the content, particularly within the dimension of viewpoint diversity, where political opinions dominated. This aligns with SOAP's design to intentionally create extreme filter bubbles, offering a controlled demonstration of how algorithmic personalization can amplify specific perspectives and limit diversity. Interestingly, while most participants found the SOAP feeds significantly different from their usual feeds, a subset reported similarities, suggesting that some users already encounter highly curated and polarized content in their personal feeds. This real-world alignment underscores the relevance of filter bubble phenomena and highlights the influence of social media algorithms in reinforcing specific viewpoints. The findings also show a notable increase in participants' awareness of viewpoint and topic diversity indicators after interacting with SOAP. Compared to pre-survey responses,

participants' post-study comments more frequently aligned with Michiels et al.'s [72] dimensions of filter bubbles. This shift indicates that exposure to SOAP-generated feeds enhanced participants' ability to recognize content biases and the lack of diversity in both topic and perspective, thus fostering a greater awareness of algorithmic influence on content curation.

Group Challenge. The diversity of ideas from the group challenge reflects a multifaceted approach to addressing filter bubbles, blending algorithmic interventions with user-focused solutions. However, it also underscores the perceived dependency on platforms and policymakers to drive these changes forward. For instance, including random–or serendipitous–content in recommended content has already been suggested by researchers [59, 69, 104], yet it is not widely implemented. The results of our group challenge suggest that it might be worthy to revisit this idea.

4.2.4 Limitations of the Workshop Study. The workshop study has several limitations that are important to keep in mind when interpreting its results, and we therefore frame it as an exploratory, proof-of-concept evaluation of SOAP's practical and educational application.

First, the participant group was relatively homogeneous, composed primarily of highly educated individuals pursuing Master's or PhD studies, which limits the generalizability of the findings to broader and more diverse populations. Second, the artificial filter bubbles generated by SOAP, while intentionally designed for controlled analysis, do not fully replicate the complexities of real-world social media environments. However, some participants noted similarities between these artificial feeds and their actual social media experiences, highlighting the practical relevance of the study. Third, the study's measurement of filter bubbles relies on participants' subjective interpretations rather than validated quantitative metrics, which may introduce variability but also offers valuable qualitative insights into user perceptions. Accordingly, the workshop results should be interpreted as evidence about perceived plausibility, usability, and pedagogical value (e.g., whether SOAP helps participants recognize personalization dynamics).

Despite these limitations, the study involved a substantial sample size of 100 participants, which strengthens the reliability of the findings. Overall, the results suggest that SOAP can support structured demonstrations of personalization dynamics and can help participants reflect on how interaction patterns shape recommendations. Complementary study designs—for example, data donation, crowd-based audits, or mixed-method field studies—can further strengthen the evidentiary basis by triangulating workshop insights with observational data from participants.

5 Conclusions

The pervasive influence of VLOPs on public discourse and individual behavior underscores the urgency for greater transparency and accountability. Driven by engagement-boosting algorithms, these platforms play a dual role: they connect people and disseminate information, but also exacerbate issues like polarization and the formation of filter bubbles. While their business models have proven immensely profitable, generating vast revenues, VLOPs often fall short in addressing their societal responsibilities and create significant negative externalities on society [45, 58, 67, 108].

To allow better oversight and scrutiny of VLOPs' recommender systems, we provide a structuring of the current regulatory landscape concerning data availability and access necessary for audits. On this basis, we provide technical tooling in the form of the SOAP framework and show that sock-puppet auditing is a viable solution to research systemic risks on VLOPs. SOAP ensures compliance with regulatory frameworks, demonstrating how technical systems can align with the legal context to promote transparency and accountability—it thereby provides a holistic framework to investigate, and potentially mitigate, the undue influence of VLOPs on democratic processes. At the same time, SOAP enables empirical research by equipping researchers to systematically audit and measure patterns associated with the emergence and persistence of systemic risks, such as

polarization through filter bubbles. To this end, we present and discuss results from a workshop with over 100 participants which showed that SOAP is an effective tool for raising awareness of systemic risks and connected phenomena such as filter bubbles. While internal or regulator-led audits may provide some insights, independent third-party auditors (e.g., academic and civil society researchers) are uniquely positioned to challenge platform narratives and uncover systemic issues that might otherwise remain hidden. By providing actionable insights into how platforms influence user experiences and shape public discourse, SOAP additionally highlights the broader societal impacts of personalized recommender systems and their role in exacerbating polarization, misinformation, and other systemic risks. SOAP is publicly accessible and can be deployed to study these phenomena at scale[34].

References

[1] Deena Abul-Fottouh, Melodie Yunju Song, and Anatoliy Gruzd. 2020. Examining algorithmic biases in YouTube's recommendations of vaccine videos. *International Journal of Medical Informatics* 140 (2020), 104175. https://doi.org/10.1016/j.ijmedinf.2020.104175

[2] Iain Akerman. 2024. Misinformation, censorship and propaganda: The information war on Gaza. https://web.archive.org/web/20250421082925/https://wired.me/business/social-media/gaza-social-media-war-palestine/ Last accessed January 13, 2026..

[3] Pascal Albisser, Keto Schumacher, Julian Schmidli, and Marina Kunz. 2025. Radikalisierung auf TikTok: Der toxische Sog der Manosphere. Schweizer Radio und Fernsehen (SRF). https://www.srf.ch/news/schweiz/radikalisierung-auf-tiktok-der-toxische-sog-der-manosphere

[4] Jef Ausloos and Michael Veale. 2021. Researching with Data Rights. *Technology and Regulation* 2020 (Jan. 2021), 136–157. https://doi.org/10.26116/techreg.2020.010

[5] Catherine Baker, Debbie Ging, and Maja Brandt Andreasen. 2024. *Recommending Toxicity: The Role of Algorithmic Recommender Functions on YouTube Shorts and TikTok in Promoting Male Supremacist Influencers.* Technical Report. DCU Anti-Bullying Centre, Dublin City University.

[6] Cameron Ballard, Ian Goldstein, Pulak Mehta, Genesis Smothers, Kejsi Take, Victoria Zhong, Rachel Greenstadt, Tobias Lauinger, and Damon McCoy. 2022. Conspiracy Brokers: Understanding the Monetization of YouTube Conspiracy Theories. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 2707–2718. https://doi.org/10.1145/3485447.3512142

[7] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. , 34 pages. https://doi.org/10.48550/arXiv.2102.04256 arXiv:2102.04256 [cs]

[8] Pablo Barberá. 2020. Social Media, Echo Chambers, and Political Polarization. In *Social Media and Democracy*, Joshua A. Tucker and Nathaniel Persily (Eds.). Cambridge University Press, Cambridge, 34–55.

[9] Judit Bayer, Natalija Bitiukova, Petra Bard, Judit Szakács, Alberto Alemanno, and Erik Uszkiewicz. 2019. *Disinformation and Propaganda – Impact on the Functioning of the Rule of Law in the EU and its Member States.* Technical Report. European Parliament, LIBE Committee, Policy Department for Citizens' Rights and Constitutional Affairs. https://doi.org/10.2139/ssrn.3409279

[10] Luka Bekavac, Kimberly Garcia, Jannis Strecker, Simon Mayer, and Aurelia Tamo-Larrieux. 2024. From Walls to Windows: Creating Transparency to Understand Filter Bubbles in Social Media. *NORMalize 2024: The Second Workshop on the Normative Design and Evaluation of Recommender Systems, co-located with the ACM Conference on Recommender Systems 2024 (RecSys 2024)* (Oct. 2024), 12. https://www.alexandria.unisg.ch/handle/20.500.14171/120987

[11] Maria Brincker. 2021. Disoriented and Alone in the "Experience Machine" – On Netflix, Shared World Deceptions and the Consequences of Deepening Algorithmic Personalization. *SATS* 22, 1 (July 2021), 75–96. https://doi.org/10.1515/sats-2021-0005

[12] Axel Bruns. 2019. After the 'APIcalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society* 22, 11 (2019), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

[13] Axel Bruns. 2019. It's not the technology, stupid: How the 'Echo Chamber' and 'Filter Bubble' metaphors have failed us. (2019). https://eprints.qut.edu.au/131675/

[14] Taina Bucher. 2018. Neither Black nor Box: (Un)knowing Algorithms. In *If...Then: Algorithmic Power and Politics.* Oxford University Press. https://doi.org/10.1093/oso/9780190493028.003.0003

[15] Ceren Budak, Brendan Nyhan, David M. Rothschild, Emily Thorson, and Duncan J. Watts. 2024. Misunderstanding the harms of online misinformation. *Nature* 630, 8015 (June 2024), 45–53. https://doi.org/10.1038/s41586-024-07417-w

[34] https://github.com/Interactions-HSG/SOAP

[16] Miriam Caroline Buiten. 2021. The Digital Services Act: From Intermediary Liability to Platform Regulation. *Social Science Research Network* (1 2021), 26. https://doi.org/10.2139/ssrn.3876328
[17] Sofia Calabrese and Orsolya Reich. 2024. Identifying, Analysing, Assessing And Mitigating Potential Negative Effects On Civic Discourse And Electoral Processes: A Minimum Menu Of Risks Very Large Online Platforms Should Take Heed of. https://www.liberties.eu/f/mpdgy5
[18] Laura Cervi and Tom Divon. 2023. Playful Activism: Memetic Performances of Palestinian Resistance in TikTok #Challenges. *Social Media + Society* 9, 1 (2023). https://doi.org/10.1177/20563051231157607
[19] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding. arXiv:2306.14924 [cs.CL]
[20] Madiha Zahrah Choksi, Marianne Aubin Le Quéré, Travis Lloyd, Ruojia Tao, James Grimmelmann, and Mor Naaman. 2024. Under the (neighbor)hood: Hyperlocal Surveillance on Nextdoor. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 771, 22 pages. https://doi.org/10.1145/3613904.3641967
[21] Pershan Claire and Amaury Lesplingart. 2024. *Full disclosure: Stress testing tech platforms' ad repositories.* Technical Report. Mozilla Foundation. https://foundation.mozilla.org/en/research/library/full-disclosure-stress-testing-tech-platforms-ad-repositories/
[22] Council of Europe. 1950. Convention for the Protection of Human Rights and Fundamental Freedoms. European Treaty Series No. 5. https://edoc.coe.int/en/european-convention-on-human-rights/5579-european-convention-on-human-rights.html
[23] Enrique Dans. 2024. It's Zuckerman vs Zuckerberg! - Enrique Dans - Medium. https://medium.com/enrique-dans/its-zuckerman-vs-zuckerberg-c4a7021e2be8. Last accesses December 13, 2024..
[24] Brittany I. Davidson, Darja Wischerath, Daniel Racek, Douglas A. Parry, Emily Godwin, Joanne Hinds, Dirk Van Der Linden, Jonathan Francis Roscoe, Laura Eeva Maria Ayravainen, and Alicia Cork. 2023. Platform-controlled social media APIs threaten open science. *Nature Human Behaviour* 7, 12 (11 2023), 2054–2057. https://doi.org/10.1038/s41562-023-01750-2
[25] Mateus Correia De Carvalho. 2024. Researcher Access to platform data and the DSA: One step forward, three steps back. https://www.techpolicy.press/researcher-access-to-platform-data-and-the-dsa-one-step-forward-three-steps-back/ Last accessed December 13, 2024.
[26] Daniel Delmonaco, Samuel Mayworm, Hibby Thach, Josh Guberman, Aurelia Augusta, and Oliver L. Haimson. 2024. "What are you doing, TikTok?": How Marginalized Social Media Users Perceive, Theorize, and "Prove" Shadowbanning. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 154 (April 2024), 39 pages. https://doi.org/10.1145/3637431
[27] Edward S Dove and Jiahong Chen. 2021. What does it mean for a data subject to make their personal data 'manifestly public'? An analysis of GDPR Article 9(2)(e). *International Data Privacy Law* 11, 2 (02 2021), 107–124. https://doi.org/10.1093/idpl/ipab005 arXiv:https://academic.oup.com/idpl/article-pdf/11/2/107/39594435/ipab005.pdf
[28] Dutch Data Protection Authority. 2024. AP: scraping bijna altijd illegaal. https://www.autoriteitpersoonsgegevens.nl/actueel/ap-scraping-bijna-altijd-illegaal Last accessed January 13,, 2026.
[29] Axel G. Ekström, Diederick C. Niehorster, and Erik J. Olsson. 2022. Self-imposed filter bubbles: Selective attention and exposure in online search. *Computers in Human Behavior Reports* 7 (2022), 100226. https://doi.org/10.1016/j.chbr.2022.100226
[30] ENISA. 2007. *Reputation-based Systems: a security analysis.* Technical Report. ENISA. https://www.enisa.europa.eu/publications/archive/reputation-based-systems-a-security-analysis/
[31] Jacob Erickson. 2024. Rethinking the Filter Bubble? Developing a Research Agenda for the Protective Filter Bubble. *Big Data & Society* 11, 1 (March 2024), 20539517241231276. https://doi.org/10.1177/20539517241231276
[32] European Court of Human Rights. 1998. Hertel v. Switzerland. 25181/94. https://hudoc.echr.coe.int/eng?i=001-59366.
[33] European Court of Human Rights. 2015. Haldimann and others v. Switzerland.
[34] European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* 59, L119 (April 2016), 1–88.
[35] European Parliament and Council of the European Union. 2019. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC. *Official Journal of the European Union* 92, L 130 (May 2019). http://data.europa.eu/eli/dir/2019/790/oj
[36] European Parliament and Council of the European Union. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act). *Official Journal of the European Union* 65, L 277 (Oct. 2022). http://data.europa.eu/eli/reg/2022/2065/oj/eng

[37] Marc Faddoul, Guillaume Chaslot, and Hany Farid. 2020. A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos. arXiv:2003.03318 [cs.CY] Computers and Society.
[38] S. Feldman. 2019. Infographic: Media Literacy Is Not a Given in Europe. Statista Daily Data. https://www.statista.com/chart/18117/media-literacy-in-europe. Last accessed January 13, 2026.
[39] K. J. Kevin Feng, Xander Koo, Lawrence Tan, Amy Bruckman, David W. McDonald, and Amy X. Zhang. 2024. Mapping the Design Space of Teachable Social Media Feed Experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 733, 20 pages. https://doi.org/10.1145/3613904.3642120
[40] Ai Forensics. 2025. Tiktok's Research API: Problems without Explanations. https://aiforensics.org/work/tk-api Last accessed January 13, 2026.
[41] Mozilla Foundation. 2024. https://www.mozillafoundation.org/en/blog/new-research-tech-platforms-data-access-initiatives-vary-widely/
[42] Geoffrey A. Fowler. 2022. Shadowbanning is real: Here's how you end up muted by social media. https://www.washingtonpost.com/technology/2022/12/27/shadowban/ Last accessed December 13, 2024.
[43] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. CollabCoder: A Lower-barrier, Rigorous Workflow for Inductive Collaborative Qualitative Analysis with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 11, 29 pages. https://doi.org/10.1145/3613904.3642002
[44] David Gilbert. 2024. TikTok Pushed Young German Voters Toward Far-Right Party. https://www.wired.com/story/tiktok-german-voters-afd Last accessed December 13, 2024.
[45] Tarleton Gillespie. 2017. Governance of and by Platforms. In *The SAGE Handbook of Social Media* (1 ed.), Jean Burgess, Alice E. Marwick, and Thomas Poell (Eds.). SAGE Publications Ltd, Los Angeles.
[46] Ayelet Gordon-Tapiero, Alexandra Wood, and Katrina Ligett. 2023. The Case for Establishing a Collective Perspective to Address the Harms of Platform Personalization. *Vanderbilt Journal of Entertainment & Technology Law* 25 (10 May 2023), 635. https://ssrn.com/abstract=4105443 Available at SSRN: https://ssrn.com/abstract=4105443.
[47] Andrew Guess, Brendan Nyhan, Benjamin Lyons, and Jason Reifler. 2018. *Why Selective Exposure to Like-Minded Political News Is Less Prevalent than You Think*. Technical Report. Knight Foundation. 26 pages.
[48] Andrew M. Guess and Benjamin A. Lyons. 2020. Misinformation, Disinformation, and Online Propaganda. In *Social Media and Democracy: The State of the Field, Prospects for Reform*, Nathaniel Persily and Joshua A. Tucker (Eds.). Cambridge University Press, 10–33.
[49] Muhammad Haroon, Magdalena Wojcieszak, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, and Zubair Shafiq. 2023. Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the National Academy of Sciences* 120, 50 (2023), e2213020120. https://doi.org/10.1073/pnas.2213020120
[50] Erik Hermann. 2022. Artificial intelligence and mass personalization of communication content—An ethical and literacy perspective. *New Media & Society* 24, 5 (2022), 1258–1277. https://doi.org/10.1177/14614448211022702
[51] Jeff Horwitz, Katherine Blunt, and Helynn Ospina for Wall Street Journal. 2023. Instagram connects vast pedophile network. https://www.wsj.com/articles/instagram-vast-pedophile-network-4ab7189 Last accessed December 13, 2024.
[52] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 48 (May 2020), 27 pages. https://doi.org/10.1145/3392854
[53] Amnesty International. 2024. Driven into Darkness: How TikTok's 'For You' Feed Encourages Self-Harm and Suicidal Ideation - Amnesty International. https://www.amnesty.org/en/documents/pol40/7350/2023/en/ Last accessed December 12, 2024.
[54] Kokil Jaidka, Subhayan Mukerjee, and Yphtach Lelkes. 2023. Silenced on social media: the gatekeeping functions of shadowbans in the American Twitterverse. *Journal of Communication* 73, 2 (1 2023), 163–178. https://doi.org/10.1093/joc/jqac050
[55] Julian Jaursch and Philipp Lorenz-Spreen. 2023. Researcher access to platform data under the DSA: Questions and answers. https://reclaimingautonomyonline.notion.site/Researcher-access-to-platform-data-under-the-DSA-Questions-and-answers-8f7390f3ae6b4aa7ad53d53158ed257c Last accessed January 13, 2026.
[56] Julian Jaursch, Jakob Ohme, and Ulrike Klinger. 2024. Enabling Research with Publicly Accessible Platform Data: Early DSA Compliance Issues and Suggestions for Improvement. *Weizenbaum Policy Paper* 9 (2024). https://doi.org/10.34669/WI.WPP/9
[57] Rishabh Kaushal, Jacob Van De Kerkhof, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. 2024. Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT*

'24). Association for Computing Machinery, New York, NY, USA, 1121–1132. https://doi.org/10.1145/3630106.3658960

[58] Batuhan Keskin. 2018. Van Dijk, Poell, and de Wall, The Platform Society: Public Values in a Connective World (2018). *Markets, Globalization & Development Review* 03 (01 2018). https://doi.org/10.23860/MGDR-2018-03-03-08

[59] Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. 2016. A Survey of Serendipity in Recommender Systems. *Knowledge-Based Systems* 111 (Nov. 2016), 180–192. https://doi.org/10.1016/j.knosys.2016.08.014

[60] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790. https://doi.org/10.1073/pnas.1320040111

[61] Jacob Kröger, Emilia Errenst, Niklas Nau, and Sanna Ojanperä. 2024. *Mitigating the Risks of Political Microtargeting – Guidance for Policymakers, Civil Society, and Development Cooperation.* Technical Report. Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), Bonn, Germany. https://www.bmz-digital.global/wp-content/uploads/2024/05/240521_Political-Microtargeting-Bericht-06.pdf.

[62] Mark Ledwich, Anna Zaitsev, and Anton Laukemper. 2022. Radical bubbles on YouTube? Revisiting algorithmic extremism with personalised recommendations. *First Monday* 27, 12 (Dec. 2022). https://doi.org/10.5210/fm.v27i12.12552

[63] Paddy Leerssen. 2020. The Soap Box as a black box: Regulating transparency in social Media recommender Systems. *European Journal of Law and Technology* 11 (2020). Issue 2. https://doi.org/10.2139/ssrn.3544009

[64] Paddy Leerssen, Amélie P. Heldt, and Matthias C. Kettemann. 2023. Scraping By? Europe's law and policy on social media research access. In *Challenges and perspectives of hate speech research*, Christian Strippel, Sünje Paasch-Colberg, Martin Emmer, and Joachim Trebbe (Eds.). Digital Communication Research, Vol. 12. Digital Communication Research, Berlin, 405–425. https://doi.org/10.48541/dcr.v12.24

[65] Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2024. Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception. arXiv:2403.14896 [cs.CY]

[66] Michele Loi. 2024. How to define platforms' systemic risks to democracy. https://algorithmwatch.org/en/making-sense-of-the-digital-services-act/ Last accessed January 13, 2026.

[67] Orla Lynskey. 2019. Grappling with "Data Power": Normative Nudges from Data Protection and Privacy. *Theoretical Inquiries in Law* 20, 1 (2019), 189–220. https://doi.org/10.1515/til-2019-0007

[68] Oliver Marsh. 2024. https://algorithmwatch.org/en/researching-systemic-risks-under-the-digital-services-act/ Last accessed January 13, 2026.

[69] Christian Matt, Alexander Benlian, Thomas Hess, and Christian Weiß. 2014. Escaping from the Filter Bubble? The Effects of Novelty and Serendipity on Users' Evaluations of Online Recommendations. *ICIS 2014 Proceedings* (Dec. 2014). https://aisel.aisnet.org/icis2014/proceedings/HumanBehavior/67

[70] Anna-Katharina Meßmer and Martin Degeling. 2023. *Auditing Recommender Systems: Putting the DSA into Practice with a Risk-Scenario-Based Approach.* Technical Report. Interface. https://interface-eu.org/publications/auditing-recommender-systems/

[71] Chang Dae Ham Michelle R. Nelson and Eric Haley. 2021. What Do We Know about Political Advertising? Not Much! Political Persuasion Knowledge and Advertising Skepticism in the United States. *Journal of Current Issues & Research in Advertising* 42, 4 (2021), 329–353. https://doi.org/10.1080/10641734.2021.1925179

[72] Lien Michiels, Jens Leysen, Annelien Smets, and Bart Goethals. 2022. What Are Filter Bubbles Really? A Review of the Conceptual and Empirical Work. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (Barcelona, Spain) *(UMAP '22 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 274–279. https://doi.org/10.1145/3511047.3538028

[73] Lien Michiels, Jorre Vannieuwenhuyze, Jens Leysen, Robin Verachtert, Annelien Smets, and Bart Goethals. 2023. How Should We Measure Filter Bubbles? A Regression Model and Evidence for Online News. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 640–651. https://doi.org/10.1145/3604915.3608805

[74] Philip-Johann Moser and Benja Zehr. 2025. AfD auf TikTok: So sieht der TikTok-Kosmos der AfD aus. https://www.zeit.de/digital/2025-02/rechts-tiktok-bundestagswahl-soziale-medien-afd

[75] OECD. 2021. *Development Co-operation Report 2021: Shaping a Just Digital Transformation.* OECD Publishing, Paris. https://doi.org/10.1787/ce08832f-en

[76] Council of Europe. 2019. Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes. *Data Protection* (Feb. 2019). https://www.coe.int/en/web/data-protection/-/declaration-by-the-committee-of-ministers-on-the-manipulative-capabilities-of-algorithmic-processes

[77] Federal Office of Justice. 1965. Copyright Act of 9 September 1965 (Federal Law Gazette I, p. 1273). as last amended by Article 25 of the Act of 23 June 2021 (Federal Law Gazette I, p. 1858). https://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html.

[78] Institute of Strategic Dialogue. 2023. Access to Social Media Data for Public Interest Research: Lessons Learnt & Recommendations for Strengthening Initiatives in the EU and Beyond - ISD. https://www.isdglobal.org/isd-publications/researcher-access-to-social-media-data-lessons-learnt-recommendations-for-strengthening-initiatives-in-the-eu-beyond/

[79] Jakob Ohme, Theo Araujo, Laura Boeschoten, Deen Freelon, Nilam Ram, Byron B. Reeves, and Thomas N. Robinson. 2024. Digital Trace Data Collection for Social Media Effects Research: APIs, Data Donation, and (Screen) Tracking. *Communication Methods and Measures* 18, 2 (April 2024), 124–141. https://doi.org/10.1080/19312458.2023.2181319

[80] Brian L. Ott. 2017. The age of Twitter: Donald J. Trump and the politics of debasement. *Critical Studies in Media Communication* 34, 1 (2017), 59–68. https://doi.org/10.1080/15295036.2016.1266686

[81] Alice Palmieri, Konrad Kollnig, and Aurelia Tamò-Larrieux. 2024. Systemic risks of dominant online platforms: A scoping review. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5002743

[82] Cecilia Panigutti, Delia Fano Yela, Lorenzo Porcaro, Astrid Bertrand, and Josep Soler Garrido. 2025. How to investigate algorithmic-driven risks in online platforms and search engines? A narrative review through the lens of the EU Digital Services Act. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 828–839. https://doi.org/10.1145/3715275.3732052

[83] Tejasi Panjiar. 2024. Instagram wants to throttle politics. https://internetfreedom.in/insta-political-content-limit/ Last accessed January 13, 2026.

[84] Kostantinos Papadamou, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Michael Sirivianos. 2022. "It Is Just a Flu": Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations. *Proceedings of the International AAAI Conference on Web and Social Media* 16, 1 (May 2022), 723–734. https://doi.org/10.1609/icwsm.v16i1.19329

[85] E. Pariser. 2011. *The Filter Bubble: What The Internet Is Hiding From You.* Penguin Books Limited. https://books.google.ch/books?id=-FWO0puw3nYC

[86] Kirsty Park, Debbie Ging, Shane Murphy, and Cian McGrath. 2023. *The Impact of the Use of Social Media on Women and Girls.* Technical Report. European Parliament, Policy Department for Citizens' Rights and Constitutional Affairs, B-1047 Brussels. This document was requested by the European Parliament's Committee on Women's Rights and Gender Equality..

[87] Kaitlyn Regehr, Caitlin Shaughnessy, Minzhu Zhao, and Nicola Shaughnessy. 2024. https://www.alignplatform.org/resources/safter-scrolling-how-algorithms-popularise-and-gamify-online-hate-and-misogyny-young

[88] Jill Rettberg. 2018. *Snapchat: Phatic Communication and Ephemeral Social Media.* University of Michigan Press.

[89] Luca Rettenberger, Markus Reischl, and Mark Schutera. 2024. Assessing Political Bias in Large Language Models. arXiv:2405.13041 [cs.CL]

[90] Urbano Reviglio and Matteo Fabbri. 2024. Navigating the Digital Services Act: Scenarios of transparency and user control in VLOPs' recommender systems. *NORMalize 2024: The Second Workshop on the Normative Design and Evaluation of Recommender Systems, co-located with the ACM Conference on Recommender Systems 2024 (RecSys 2024), Bari, Italy.* (2024), 9.

[91] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 131–141. https://doi.org/10.1145/3351095.3372879

[92] Jan Ole Rixen, Luca-Maxim Meinhardt, Michael Glöckler, Marius-Lukas Ziegenbein, Anna Schlothauer, Mark Colley, Enrico Rukzio, and Jan Gugenheimer. 2023. The Loop and Reasons to Break It: Investigating Infinite Scrolling Behaviour in Social Media Applications and Reasons to Stop. *Proc. ACM Hum.-Comput. Interact.* 7, MHCI, Article 228 (Sept. 2023), 22 pages. https://doi.org/10.1145/3604275

[93] David Rozado. 2024. The Political Preferences of LLMs. arXiv:2402.01789 [cs.CY] Computers and Society.

[94] Santiago Sordo Ruz, Martin Degeling, and Kathy Meßmer. 2023. The Research API falls woefully short. https://tiktok-audit.com/blog/2023/the-TikTok-research-API-falls-woefully-short/ Last accessed January 13, 2026.

[95] Matthew Sadiku, Tolulope Joshua Ashaolu, Abayomi Ajayi-Majebi, and Sarhan Musa. 2021. Artificial Intelligence in Social Media. *International Journal Of Scientific Advances* 2 (01 2021). https://doi.org/10.51542/ijscia.v2i1.4

[96] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In *"Data and Discrimination: Converting Critical Concerns into Productive Inquiry" Co-Located with the 64th Annual Meeting of the International Communication Association.* 23.

[97] Mia Sato. 2024. Pro-Harris TikTok Felt Safe in an Algorithmic Bubble — until Election Day. https://www.theverge.com/2024/11/14/24295814/kamala-harris-tiktok-filter-bubble-donald-trump-algorithm Last accessed January 13, 2026.

[98] Rebecca Sawyer and Guo-Ming Chen. 2012. The impact of social media on intercultural adaptation. https://digitalcommons.uri.edu/com_facpubs/15/

[99] Sam Schechner, Rob Barry, Georgia Wells, Jason French, Brian Whitton, and Kara Dapena. 2024. How TikTok Brings War Home to Your Child. https://www.tovima.com/wsj/how-tiktok-brings-war-home-to-your-child/ Last accessed January 13, 2026.

[100] Christophe Olivier Schneble, Bernice Simone Elger, and David Shaw. 2018. The Cambridge Analytica affair and Internet-mediated research. *EMBO reports* 19, 8 (2018), e46579. https://doi.org/10.15252/embr.201846579

[101] Andrea Schneiker, Magnus Dau, Jutta Joachim, Marlen Martin, and Henriette Lange. 2018. How to Analyze Social Media? Assessing the Promise of Mixed-Methods Designs for Studying the Twitter Feeds of PMSCs. *International Studies Perspectives* 20, 2 (11 2018), 188–200. https://doi.org/10.1093/isp/eky013

[102] Jaime E Settle. 2018. *Frenemies: How social media polarizes America.* Cambridge University Press.

[103] Toby Shevlane and Allan Dafoe. 2020. The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 173–179. https://doi.org/10.1145/3375627.3375815

[104] Annelien Smets, Lien Michiels, Toine Bogers, and Lennart Björneborn. 2022. Serendipity in Recommender Systems Beyond the Algorithm: A Feature Repository and Experimental Design. *CEUR Workshop Proceedings* 3222, 46–66.

[105] Sophie. 2024. Mozilla commissions CheckFirst to conduct a stress test of Ad Repositories. https://checkfirst.network/mozilla-commissions-checkfirst-to-conduct-a-stress-test-of-ad-repositories/

[106] Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2023. The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 17, 4 (Sep. 2023), Article 8. https://doi.org/10.5817/CP2023-4-8

[107] Larissa Spinelli and Mark Crovella. 2020. How YouTube Leads Privacy-Seeking Users Away from Reliable Information. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 244–251. https://doi.org/10.1145/3386392.3399566

[108] Nick Srnicek. 2017. *Platform Capitalism.* Polity, Cambridge.

[109] Chris Stokel-Walker. 2023. Twitter's $42,000-per-Month API prices out nearly everyone. *WIRED* (March 2023). https://www.wired.com/story/twitter-data-api-prices-out-nearly-everyone/ Last accessed January 13, 2026.

[110] Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Chloe Bakalar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, Connie Moon Sehat, Sara Johansen, Lianne Kerlin, David Vickrey, Spandana Singh, Sanne Vrijenhoek, Amy Zhang, McKane Andrus, Natali Helberger, Polina Proutskova, Tanushree Mitra, and Nina Vasan. 2024. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *ACM Transactions on Recommender Systems* 2, 3 (June 2024), 1–57. https://doi.org/10.1145/3632297

[111] Laura Sudulich, Matthew Wall, Rachel Gibson, Marta Cantijoch, and Stephen Ward. 2014. *Introduction: The Importance of Method in the Study of the 'Political Internet'.* Palgrave Macmillan UK, London, 1–21. https://doi.org/10.1057/9781137276773_1

[112] Mubashir Sultan, Christin Scholz, and Wouter van den Bos. 2023. Leaving Traces behind: Using Social Media Digital Trace Data to Study Adolescent Wellbeing. *Computers in Human Behavior Reports* 10 (May 2023), 100281. https://doi.org/10.1016/j.chbr.2023.100281

[113] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic Biases in LLM Simulations of Debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 251–267. https://doi.org/10.18653/v1/2024.emnlp-main.16

[114] Ludovic Terren and Rosa Borge-Bravo. 2021. Echo Chambers on Social Media: A Systematic Review of the Literature. *Review of Communication Research* 9 (March 2021), 99–118. https://doi.org/10.12840/ISSN.2255-4165.028

[115] Jasper Tjaden, Johannes Wolfgram, Aaron Philipp, Sarah Weissmann, Licia Bobzien, Ulrich Kohler, and Roland Verwiebe. 2024. Automated election audits - Analyzing Exposure to Political Content on Social Media with a Case Study of TikTok in Germany's 2024 Regional Elections. https://doi.org/10.31219/osf.io/qwrjh

[116] Matus Tomlein, Branislav Pecher, Jakub Simko, Ivan Srba, Robert Moro, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, and Maria Bielikova. 2021. An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3460231.3474241

[117] Amaury Trujillo, Tiziano Fagni, and Stefano Cresci. 2023. The DSA Transparency Database: Auditing Self-reported Moderation Actions by Social Media. *arXiv (Cornell University)* (12 2023). https://doi.org/10.48550/arxiv.2312.10269

[118] Aleksandra Urman, Mykola Makhortykh, and Aniko Hannak. 2025. WEIRD Audits? Research Trends, Linguistic and Geographical Disparities in the Algorithm Audits of Online Platforms - A Systematic Literature Review. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing

Machinery, New York, NY, USA, 375–390. https://doi.org/10.1145/3715275.3732026

[119] Irene I. van Driel, J. Loes Pouwels Anastasia Giachanou, Laura Boeschoten, Ine Beyens, and Patti M. Valkenburg. 2022. Promises and Pitfalls of Social Media Data Donations. *Communication Methods and Measures* 16, 4 (2022), 266–282. https://doi.org/10.1080/19312458.2022.2109608

[120] Tim Verbeij, Ine Beyens, Damian Trilling, and Patti M. Valkenburg. 2024. Happiness and Sadness in Adolescents' Instagram Direct Messaging: A Neural Topic Modeling Approach. *Social Media + Society* 10, 1 (2024), 20563051241229655. https://doi.org/10.1177/20563051241229655

[121] Matthijs Warrens. 2015. Five Ways to Look at Cohen's Kappa. *Journal of Psychology & Psychotherapy* 05 (01 2015). https://doi.org/10.4172/2161-0487.1000197

[122] Joe Whittaker, Seán Looney, Alastair Reed, and Fabio Votta. 2021. Recommender systems and the amplification of extremist content. https://policyreview.info/articles/analysis/recommender-systems-and-amplification-extremist-content

[123] Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23 Companion)*. ACM, New York, NY, USA, 75–78. https://doi.org/10.1145/3581754.3584136

[124] Vithya Yogarajan, Gillian Dobbie, Te Taka Keegan, and Rostam J Neuwirth. 2023. Tackling bias in pre-trained language models: Current trends and under-represented societies. arXiv:2312.01509

[125] Rasha Younes. 2023. Meta's Broken Promises. https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and Last accessed January 13, 2026..

[126] Savvas Zannettou, Olivia-Nemes Nemeth, Oshrat Ayalon, Angelica Goetzen, Krishna P. Gummadi, Elissa M. Redmiles, and Franziska Roesner. 2023. Leveraging Rights of Data Subjects for Social Media Analysis: Studying TikTok via Data Donations. https://doi.org/10.48550/arXiv.2301.04945 arXiv:2301.04945 [cs]

[127] Brahim Zarouali, Sophie C. Boerman, and Claes H. de Vreese. 2021. Is this recommended by an algorithm? The development and validation of the algorithmic media content awareness scale (AMCA-scale). *Telematics and Informatics* 62 (2021), 101607. https://doi.org/10.1016/j.tele.2021.101607

A Safety Filter Response

```
ERROR:root:Unexpected error: Response has no candidates (and thus no text
    ). The response is likely blocked by the safety filters.
Response:
{
  "prompt_feedback": {
    "block_reason": "PROHIBITED_CONTENT"
  },
  "usage_metadata": {
    "prompt_token_count": 32733,
    "total_token_count": 32733
  }
}
```

B Primer Prompts

B.1 Aviation Filter Bubble

```
You are analyzing Social Media videos about aviation and aircrafts.
Post Description: {post_text}
Username: {creator_id}
Please answer the following Questions:
- How much is the video about aviation/aircrafts/flying on a binary scale
 of 0 to 1? (0 being not at all about aviation and 1 being about aviation
)
- What is your reasoning for the score?
Either the video itself or the description should contain aviation/
aircrafts/flying related content.
Give your answer precisely in the following format:
"Score:Reasoning". For example, "1:The video is entirely about different
types of aircraft and their mechanisms. Do not say Score or Reasoning in
the answer.
```

B.2 Kitten Filter Bubble

```
"You are analyzing Social Media videos that feature cute and adorable
kittens.
Post Description: {post_text}
Username: {creator_id}
Please answer the following questions:
- Does the video feature or relate to cute and adorable kittens? (0 for
No and 1 for Yes)
- What is your reasoning for the score?
Consider the following topics as examples and rate higher if the video
prominently features, but limietd to, these topics:
- Kitten antics (e.g., playful behavior, funny moments)
- Kitten care (e.g., grooming tips, health advice, feeding information)
- Kitten milestones (e.g., first steps, learning to purr, first time
playing with toys)
```

```
- Kitten adoption stories (e.g., rescue tales, adoption success stories)
- Kitten interactions with humans and other animals (e.g., cuddling with
owners, playing with other pets)
- Kitten habitats (e.g., cozy beds, playful environments, safe outdoor
explorations)
For example:
- A video showing kittens playing with each other should be rated 1.
- A video providing tips on how to care for a new kitten should be rated
1.
- A video showing a kitten being adopted into a loving home should be
rated 1.
- A video presenting a kitten's first time exploring a new environment
should be rated 1.
- A video exploring different types of kitten toys and how kittens
interact with them should be rated 1.
Give your answer precisely in the following format:
""Score:Reasoning"". For example, ""1:The video is entirely about kittens
 playing and showing their adorable antics."" Do not say 'Score' or '
Reasoning' in the answer."
```

B.3 Palestine/Israel Bubble

```
You are analyzing Social Media videos about the Palestine/Israel conflict
Post Description: {post_text}
Username: {creator_id}
Please answer the following question:
- Is the video about the Palestine/Israel conflict? (0 being not at all
about the conflict and 1 being about the conflict. Only provide a binary
score of 0 or 1)
Either the video itself or the description should contain content related
 to the Palestine/Israel conflict. It is also sufficient if the video is
about events and issues surrounding the conflict, like providing aid to
Gaza.
Give your answer precisely in the following format:
"Score:Reasoning". For example, "1:The video is about the events and
issues surrounding the Palestine/Israel conflict. Do not say 'Score' or '
Reasoning' in your answer."
```

B.4 US-Election Bubble

```
You are analyzing Social Media videos related to the US election and US
politics from the perspective of a voting US American.
Post Description: {post_text}
Username: {creator_id}
Please answer the following questions:
- On a scale of 0 to 5, how much is the video related to the US election,
 US politics, or the US voters in general? (0 being not at all related
and 5 being entirely focused on this theme)
```

```
- What is your reasoning for the score?
If the post is related to the US election or US politics, rate it higher.
 Videos that feature US politicians, political debates, or legislative
updates should also be rated high.
Give your answer precisely in the following format:
"Score:Reasoning". For example, "5:The video is entirely about positive
aspects of left-leaning US politics." Do not say 'Score' or 'Reasoning'
in the answer."
```

C Bot accounts used for Workshop Filter Bubbles

Table 4. Usernames of the bot accounts on Instagram used for filter bubble exploration in the Workshop with the usernames of the accounts they followed.

Username	Usernames of Followed Accounts
smblackbox3	none
smblackbox6	americanvalues2024, allinwithchris, newsnationnow, realdonaldtrump, gop
Will_Byers44	thedailybeast, huffpost, msnbc, slate, realdonaldtrump, joebiden, nytopinion
Dustin_Henderson44	thefivefnc, usairforce, usarmy, realdonaldtrump, foxnews, fncoriginals, teamtrump, seanhannity, realamericasvoice, therudygiuliani, newsmax
agila.yaz44	none
peterpan6048	foxnation, repmattgaetz, realmarjoriegreene, realdonaldtrump, whitehouse, rondesantis, melaniatrumpworld, erictrump, jessewatters, senatorvance, jdvance, newsmax, officialbenshapiro, melaniatrump, teamtrump, donaldjtrumpjr
Max_Mayfield44	none
bugo.kul44	none
redload44	demgovs, senatedems, dscc, facethenation, senbooker, vote4democrats, scdemparty, americasvoice, rollcall, positive, thedemocrats, kamalaharris
tanyablak44	seanhannity, senatorvance, donaldjtrumpjr, realdonaldtrump, joebiden, kamalahq, potus, timwalz, kamalaharris

D Surveys administered during the Workshop

D.1 Pre-Survey

(1) I consent to my data being collected, stored, and used for research purposes related to this questionnaire, specifically for research on filter bubbles and social media algorithms. I understand that all data will be collected anonymously and handled in accordance with data protection regulations. (Yes, No)
(2) I want to be informed via email about the study's outcome.
(3) How old are you?
(4) What is your Gender? (Woman, Man, Non-binary, Prefer not to say, Self-described)
(5) In what programm are you enrolled at the University of St.Gallen?

(6) AMCA scale (adapted from [127]; on a 5-point Likert scale: 1 = "not at all aware", 5 = "Completely aware")
- FIL1 Algorithms are used to recommend posts/content to me on Social Media
- FIL2 Algorithms are used to prioritize certain posts/content above others
- FIL3 Algorithms are used to tailor certain posts/content to me on Social Media
- FIL4 Algorithms are used to show someone else see different posts/content than I get to see on Social Media
- ADM1 Algorithms are used to show me posts/content on Social Media based on automated decisions
- ADM2 Algorithms do not require human judgments in deciding which posts/content to show me on Social Media
- ADM3 Algorithms make automated decisions on what posts/content I get to see on Social Media
- HAI1 The posts/content that algorithms recommend to me on Social Media depend on my online behavior on that platform
- HAI2 The posts/content that algorithms recommend to me on Social Media depend on my online behavioral data
- HAI3 The posts/content that algorithms recommend to me on Social Media depend on the data that I make available online
- ETH1 It is not always transparent why algorithms decide to show me certain posts/content on Social Media
- ETH2 The posts/content that algorithms recommend to me on Social Media can be subjected to human biases such as prejudices and stereotypes
- ETH3 Algorithms use my personal data to recommend certain posts/content on Social Media, and this has consequences for my online privacy

(7) How would you describe the tone and content of your typical social media feed? Partisan = strongly supporting a person, principle, or political party, often without considering or judging the matter very carefully. Neutral = unbiased and balanced, presenting information or content without favoring any side, person, principle, or political party. The content aims to be fair and does not promote a specific viewpoint or agenda. (5-point Likert scale: 1 = "very neutral", 5 = "very partisan")

(8) How likely do you think social media feeds could influence your personal political opinions or beliefs over time? (5-point Likert scale: 1 = "Not likely at all", 5 = "Very likely")

(9) How credible do you find the information presented in your typical social media feed? (5-point Likert scale: 1 = "Not credible at all", 5 = "Very credible")

(10) What do you think are clear signs or indicators that suggest content is heavily curated or biased in a social media feed? (Free text entry)

D.2 Post-Survey

(1) I consent to my data being collected, stored, and used for research purposes related to this questionnaire, specifically for research on filter bubbles and social media algorithms. I understand that all data will be collected anonymously and handled in accordance with data protection regulations. (Yes, No)

(2) Which Group number were you assigned to?

(3) AMCA scale (see above in D.1)

(4) How would you describe the tone and content of the feed you analyzed? Partisan = strongly supporting a person, principle, or political party, often without considering or judging the matter very carefully. Neutral = unbiased and balanced, presenting information or content without favoring any side, person, principle, or political party. The content aims to be fair

and does not promote a specific viewpoint or agenda. (5-point Likert scale: 1 = "very neutral", 5 = "very partisan")

(5) How likely is it that consistently being exposed to this feed would influence your personal political opinions or beliefs over time? (5-point Likert scale: 1 = "Not likely at all", 5 = "Very likely")

(6) How credible do you find the information presented in the feed? (5-point Likert scale: 1 = "Not credible at all", 5 = "Very credible")

(7) Did you notice any clear signs or indicators that suggested the content was part of a filter bubble? (Free text entry)

(8) What Social Media Platforms do you use personally? (Instagram, Facebook, Snapchat, Twitter/X, TikTok, Mastodon, LinkedIn, None, Other)

(9) Have you seen political posts/content in your feed recently? (Yes, No, Unsure)

(10) How does the content in the filter bubble feed compare to what you typically see in your own social media feeds? (Free text entry)

(11) Would you say your feed has been or is in a filter bubble currently? (Of any topic, e.g politics, sport, education) (Yes, No)

(12) If Yes:

(12a) On which platform did you encounter the filter bubble? (Instagram, Facebook, Snapchat, Twitter/X, TikTok, Mastodon, LinkedIn, None, Other; multiple possible)

(12b) What was/is the topic of the filter bubble? (Free text entry)

(12c) How often do you believe you encounter filter bubbles in your own social media use? (Very frequently, Occasionally, Rarely, Never)

(13) Would you be willing to donate your public Explore page Social Media Feed for further analysis? We will contact you in this case for a follow up study (Yes, No)

(14) Please provide any further comments or feedback below (Free text entry)

E Additional Data from the Workshop

Table 5. Detailed demographics of the workshop participants who filled both surveys (N=100).

		Count
Gender	Woman	44
	Man	56
Age in years	21	6
	22	25
	23	28
	24	17
	25	13
	26	2
	27	5
	28	1
	30	1
	32	1
	36	1
Pursued Degree	Master	96
	PhD	4
Program (Master)	Accounting and Corporate Finance	36
	Banking and Finance	11
	Marketing Management	11
	General Management	9
	International Affairs and Governance	7
	Strategy and International Management	4
	Business Innovation	4
	Economics	4
	Computer Science	2
	Management, Organisation and Culture	2
	Quantitative Economics and Finance	2
	Law and Economics	2
	Law	1
	International Law	1
Program (PhD)	Computer Science	1
	Economics and Econometrics	1
	Finance	1
	Law	1

Table 6. Platform usage by the workshop participants (N=100; multiple selections possible), number of filter bubbles encountered per platform, and percentage of these encounters relative to the platform usage numbers.

		Count	Filter Bubbles Encountered	Percentage (rounded)
Platform	Instagram	92	69	75%
	LinkedIn	92	6	6%
	Snapchat	63	3	4%
	TikTok	40	28	70%
	Facebook	34	8	23%
	Twitter/X	17	14	82%
	Reddit	3	1	33%
	BeReal	1	0	
	Discord	1	0	
	Rumble	1	0	
	Twitch	1	0	
	WIWI treff	1	0	
	YouTube	1	1	100%
Filter Bubble encountered recently	Yes	79		
	No	21		
Frequency of Filter Bubble Encounters (N=79)	Very frequently	60		
	Occasionally	17		
	Rarely	2		

Table 7. A comparison of user perceptions of their own feeds compared to the SOAP-created feeds with the results of Wilcoxon signed-rank tests and the rank-biserial correlation r indicating the effect size.

		Pre: Mean (SD)	Post: Mean (SD)	W	p-value	r
Q1	How would you describe the tone and content of your typical social media feed?	2.83 (1.04)				
	How would you describe the tone and content of the feed you analyzed? (1= "very neutral", 5 = "very partisan")		4.09 (0.93)	191	<0.001	-0.972
Q2	How likely do you think social media feeds could influence your personal political opinions or beliefs over time?	3.32 (1.06)				
	How likely is it that consistently being exposed to this feed would influence your personal political opinions or beliefs over time? (1= "not likely at all", 5 = "very likely")		3.71 (1.17)	782	0.003	-0.854
Q3	How credible do you find the information presented in your typical social media feed?	2.83 (0.83)				
	How credible do you find the information presented in the feed? (1= "not credible at all", 5 = "very credible")		2.31 (1.13)	642	<0.001	-0.882

F Code Repository

- All code and instructions can be found on https://github.com/Interactions-HSG/SOAP.
- The full social media data used for deductive coding evaluation and the exploration of filter bubbles is stored on university servers and available upon request.