



International Journal of Journalism and Mass Communication Research

Journal homepage: <https://trpub.org/journal.php?id=gAxgAuOMGx>

TRPub

IJJMCR

International Journal of
Journalism and Mass
Communication
Research

Vol. 1 Issue 1 2026

Methodology Article

Open Access

Auditing Short-Video Recommendation Feeds With Sock-Puppet Accounts: A Replicable Design

Liam Beckford^{1*}, Beatriz Subramaniam², Sebastien Ellery³

¹Department of Social Sciences, Nejd Institute of Social Sciences, Romania

²Department of Social Sciences, Orinoco University Centre for Social Sciences, Kyrgyzstan

³Department of Social Sciences, Elbe University Centre for Social Sciences, Kyrgyzstan

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijjmc.2026.v1.i1.pp19-24](https://doi.org/10.62715/ijjmc.2026.v1.i1.pp19-24)

Received: 05 February 2026

Accepted: 03 April 2026

Published: 01 June 2026

Abstract

Short-video platforms deliver most of what their users see through a recommendation feed that requires no search, no subscription and often no deliberate act beyond continuing to watch. Because the signal the system learns from is largely implicit, auditing such feeds with automated accounts raises design problems that earlier audits of search engines and long-form video did not face. This article proposes a replicable design for sock-puppet audits of short-video feeds. It reviews the audit tradition and the published studies that have used scripted accounts, identifies the features of short-video recommendation that matter for design, and sets out an eight-step protocol: defining the question and outcome, specifying personas with matched controls, building a documented technical environment, scripting viewing behaviour, piloting and registering before running accounts in parallel, capturing and archiving the feed, annotating content with a tested codebook, and analysing differences against a measured noise floor. Threats to validity are discussed, in particular the gap between scripted and human behaviour, platform detection of automation and the instability of results over time. Legal and ethical conditions are outlined. The protocol is a proposal built from existing practice and has not been run or validated for this article; what a validation would need to show is stated.

Keywords: *Algorithm audit, Sock-puppet accounts, Recommender systems, Short-video platforms, Personalisation, Research replicability*

© 2026 Liam Beckford et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

Public argument about what recommendation systems show people has outrun the means of finding out. Platform operators hold the logs. Outside researchers can ask users to donate their data, can observe their own feeds, or can construct accounts that behave in a controlled way and record what those accounts are served. The last approach, the sock-puppet audit, is one of the designs set out in the foundational typology of Sandvig et al. (2014), and it has become a standard tool for studying personalisation (Bandy, 2021; Metaxa et al., 2021).

* Corresponding author: Liam Beckford, Department of Social Sciences, Nejd Institute of Social Sciences, Romania. E-mail: liam.beckford@nejd.edu

© 2026 Liam Beckford et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Most of the method's conventions were developed on search engines and on long-form video, where a user expresses interest by typing a query, clicking a result or choosing a video. A short-video feed works differently. Content plays automatically in a single stream, and the system infers interest chiefly from how long each item is watched before the user swipes on. An audit therefore has to manipulate dwell time with care, and many more items pass in a session, so the volume of material to be classified is large. Published audits of these feeds exist, but each has made its own choices about accounts, scripts and controls, and the choices are not always reported in enough detail to repeat.

This article proposes a design that can be repeated. It is a protocol, with reasons, for researchers in journalism and communication who want to ask what a short-video feed serves to accounts with different interests in news and public affairs. The protocol has not been executed for this article. It is assembled from published practice and stated as a proposal to be tested.

2. Background

The design draws on the general logic of algorithm audits, on the experience of earlier sock-puppet studies, and on what is publicly known about how short-video recommendation behaves.

2.1. The audit tradition

[Sandvig et al. \(2014\)](#) borrowed the audit from field studies of discrimination, in which matched testers who differ in one characteristic approach the same landlord or employer. Applied to software, the matched testers are accounts and the characteristic is some input to the system. The method is experimental in form: because the researcher controls the input, differences in output can be attributed to it, provided other sources of variation are held constant or measured. [Hannak et al. \(2013\)](#) showed how much care that proviso demands. Studying web search, they found that results differed between identical accounts queried at the same moment, because of load balancing, experiments run by the platform and carry-over between queries, and they introduced paired control accounts to measure this noise so that it would not be mistaken for personalisation. [Haim et al. \(2018\)](#) used scripted agents in the same spirit to test for filter bubbles in a news aggregator.

2.2. Sock-puppet studies of video recommendation

On long-form video, [Hussein et al. \(2020\)](#) built accounts with defined watch histories and measured the misinformation subsequently recommended to them, and [Tomlein et al. \(2021\)](#) tested whether accounts that had been drawn into such recommendations could be drawn out again by watching corrective material. [Haroon et al. \(2023\)](#) ran the approach at very large scale to examine whether recommendations matched and intensified the political leaning of an account's history. For short video, [Boeker and Urman \(2022\)](#) used automated accounts to test, one factor at a time, how location, language, following, liking and watch duration affected the feed, and [Vombatkere et al. \(2024\)](#) combined automated accounts with data from real users to separate recommendations that exploit known interests from those that explore new ones.

A parallel literature cautions against reading too much into puppets. Studies that observe real users find that exposure to extreme channels is concentrated among people who already seek them out through subscriptions and external links ([Chen et al., 2023](#)), and [Hosseinmardi et al. \(2024\)](#) used bots that replay real viewing histories to show that following recommendations blindly can lead to different content from what the same users chose for themselves. A sock puppet measures what the system offers to a specified pattern of behaviour. It does not measure what people consume.

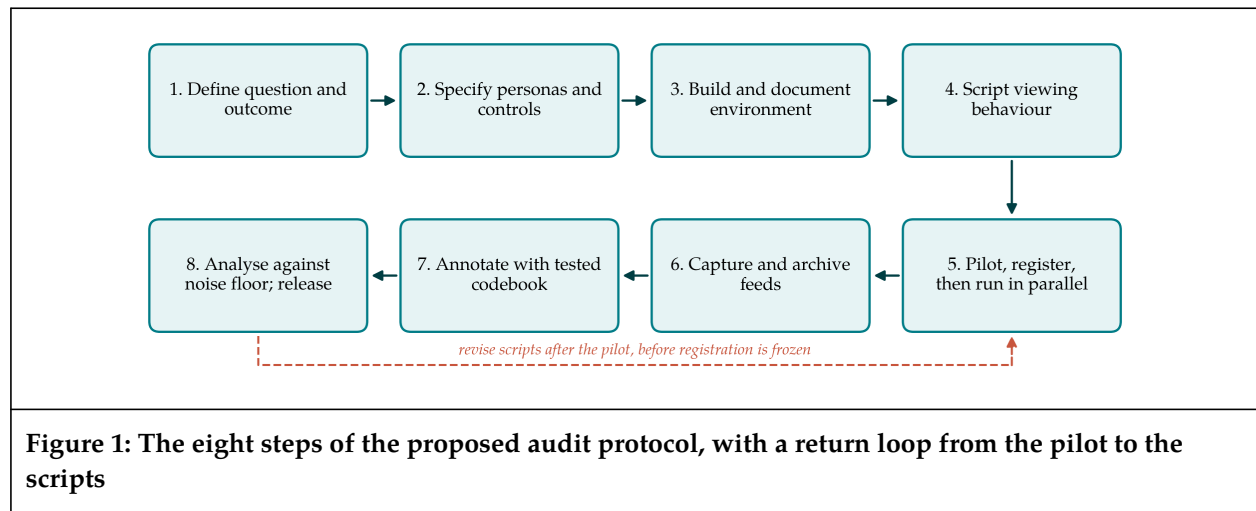
2.3. What is distinctive about short-video feeds

Industrial recommenders are generally built in two stages, a broad retrieval of candidates followed by a ranking model trained on engagement signals such as expected watch time ([Covington et al., 2016](#)). On a short-video feed the engagement signal is abundant and mostly implicit: each item yields a completion or an early swipe. The audit evidence indicates that watch duration, alongside following and liking, shapes what an account is shown ([Boeker and Urman, 2022](#)). Three design consequences follow. Dwell time is the primary treatment variable and must be scripted precisely. A new account receives a feed shaped by

its declared location, language and device before it has done anything, so these must be fixed and reported. And the feed adapts within a session, so the order in which items arrive is itself data.

3. The Proposed Protocol

The protocol has eight steps. Figure 1 shows their order and Table 1 states, for each, the main decision to be taken and what must be reported for another team to repeat the work.



Step 1 fixes the question and the outcome. A question such as whether accounts that linger on election content are subsequently served a larger share of content from unverified sources than matched controls names a treatment, a comparison and a measurable property of feed items. Vague aims, for example to see what the algorithm does, cannot be audited. Step 2 specifies the personas. Each treatment persona is defined by the content categories it dwells on, and each is matched with at least two control accounts that share its location, language, device profile and creation time: one idle control that watches every item for the same short duration, and one duplicate of the treatment persona. The duplicate supplies the noise floor in the manner of [Hannak et al. \(2013\)](#).

Table 1: Steps of the proposed protocol, the main design decision at each, and the minimum that should be reported

Step	Main design decision	Minimum to report
1. Question and outcome	Treatment, comparison and the feed property measured	Hypotheses as registered
2. Personas and controls	Interest profiles; idle and duplicate controls	Number of accounts per condition and why
3. Environment	Devices or emulators, network location, app version	Full configuration and account creation steps
4. Behaviour scripts	Rule for classifying each item; dwell times; other actions	Script code and classification rule
5. Pilot and run	Session length, schedule, number of replicates	Dates, times, deviations from plan
6. Capture	What is logged for each item and how it is stored	Fields captured; archive location
7. Annotation	Codebook, coder training, reliability test	Codebook and agreement coefficients
8. Analysis and release	Estimand, comparison with duplicate controls	Analysis code, item identifiers, limits

Step 3 builds the environment. Short-video platforms are used chiefly through mobile applications, and a browser session may not receive the same feed. The protocol therefore prefers physical handsets or emulated devices running the application, each with a fresh installation, no contact list and a network location that matches the persona's declared country. The application version, operating system, location method and every step of account creation are recorded. Step 4 writes the behaviour scripts. A script needs a rule for deciding, as each item arrives, whether it belongs to the persona's interest category; the rule can use captions, hashtags and sounds, and it should be simple enough to publish. Items that match are watched to the end, others are skipped after a fixed short interval, and small random variation is added to both. Likes, follows and searches are either excluded or introduced as separate conditions, so that their effects are not confounded with dwell time.

Step 5 pilots, registers and runs. A pilot with a few accounts exposes practical failures: login challenges, interstitial screens, scripts that misclassify. After the pilot the design is registered with its hypotheses and analysis plan, in keeping with the transparency standards proposed for empirical research generally (Nosek et al., 2015). All accounts in a replicate then run at the same times of day over the same days, since feeds respond to current events. Step 6 captures the feed: for every item, its identifier, position in the session, timestamp, creator, caption, hashtags, sound, and visible engagement counts, with the video file or a screenshot archived where storage and law allow, because items are frequently removed.

Step 7 annotates. Whatever property the outcome depends on, such as topic, source type or accuracy rating, is coded from a written codebook by at least two trained coders working independently on a common subset, and agreement is reported with a chance-corrected coefficient such as Krippendorff's alpha (Krippendorff, 2018). Automated classifiers may be used to extend the coding once they have been checked against the human-coded subset. Step 8 analyses and releases. The quantity of interest is the difference between treatment and idle control in the share of feed items with the coded property, and it is interpreted only in relation to the difference observed between duplicate accounts. Scripts, configuration, codebook, item identifiers and analysis code are deposited so that the audit can be rerun.

4. Threats to Validity

Four threats deserve explicit treatment in any report. The first is ecological: a puppet that watches one category with perfect consistency is unlike any person, and results describe the system's response to an extreme input (Chen et al., 2023; Hosseinmardi et al., 2024). Personas can be made less artificial by mixing interests, at the cost of a weaker treatment. The second is detection. Platforms act against automated accounts, and an account that has been flagged may be served a degraded or atypical feed without notice; sudden changes in a puppet's feed relative to its duplicate are a warning sign. The third is temporal. Recommendation models, content supply and moderation rules change continually, so an audit is a dated observation, and the systematic review by Bandy (2021) notes how few audits are repeated. The fourth is measurement: an outcome that depends on unreliable coding cannot support a conclusion however clean the experimental design.

5. Legal and Ethical Conditions

Creating accounts under invented identities and collecting data by automated means are commonly prohibited by platform terms of service, although such terms are often ambiguous and unevenly enforced (Fiesler et al., 2020). Researchers should take advice on the law of their jurisdiction before starting. In the European Union the Digital Services Act creates routes for vetted researchers to obtain data from very large platforms and addresses research use of publicly accessible data (European Union, 2022), which may in time reduce the need for puppets but does not replace the independent check that an external audit provides. Ethical review should cover the interests of the creators whose videos are collected, who have not consented to study. The guidelines of the Association of Internet Researchers recommend proportionate collection, secure storage and care in republishing identifiable material (franzke et al., 2020). Puppets should not post, comment or message, and should be few enough that they do not distort the engagement counts of small creators.

6. Limitations and What Remains to Be Shown

The protocol is proposed and unvalidated. A validation would need to establish at least three things. Independent teams following the protocol with the same registered design at the same time should obtain compatible estimates, which would show that the reporting requirements are sufficient. Differences between duplicate accounts should be small relative to differences between treatment and control, which would show that the noise floor is usable in this setting. And feeds served to emulated devices should be compared with feeds on physical handsets, since the equivalence is assumed here and not demonstrated. The protocol also leaves open how many accounts and sessions are enough. That depends on the variability of the feed, which only a pilot on the platform in question can reveal.

7. Conclusion

Sock-puppet audits remain one of the few ways for outsiders to test claims about what recommendation feeds serve. On short-video platforms the method must be adapted to a system that learns mainly from watching time, shapes a feed from the first moment of an account's life and changes quickly. The design proposed here responds with matched idle and duplicate controls, precisely scripted and published dwell rules, a documented device environment, registration after piloting, reliability-tested annotation and full release of materials. Its worth will be settled by teams who run it and report where it fails.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Bandy, J. (2021). [Problematic machine behavior: A systematic literature review of algorithm audits](#). *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): Article 74. <https://doi.org/10.1145/3449148>
- Boeker, M. and Urman, A. (2022). [An empirical investigation of personalization factors on TikTok](#). *Proceedings of the ACM Web Conference 2022*, 2298-2309.
- Chen, A.Y., Nyhan, B., Reifler, J., Robertson, R.E. and Wilson, C. (2023). [Subscriptions and external links help drive resentful users to alternative and extremist YouTube channels](#). *Science Advances*, 9(35): eadd8080. <https://doi.org/10.1126/sciadv.add8080>
- Covington, P., Adams, J. and Sargin, E. (2016). [Deep neural networks for YouTube recommendations](#). *Proceedings of the 10th ACM Conference on Recommender Systems*, 191-198. <https://doi.org/10.1145/2959100.2959190>
- European Union (2022). [Regulation \(EU\) 2022/2065 of the European Parliament and of the Council on a Single Market for Digital Services \(Digital Services Act\)](#). Official Journal of the European Union, L 277, 27 October 2022.
- Fiesler, C., Beard, N. and Keegan, B.C. (2020). [No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 14: 187-196.
- franzke, a.s., Bechmann, A., Zimmer, M., Ess, C. and the Association of Internet Researchers (2020). [Internet Research: Ethical Guidelines 3.0](#). Association of Internet Researchers.
- Haim, M., Graefe, A. and Brosius, H.-B. (2018). [Burst of the filter bubble? Effects of personalization on the diversity of Google News](#). *Digital Journalism*, 6(3): 330-343. <https://doi.org/10.1080/21670811.2017.1338145>
- Hannak, A., Sapiezynski, P., Kakhki, A.M., Krishnamurthy, B., Lazer, D., Mislove, A. and Wilson, C. (2013). [Measuring personalization of web search](#). *Proceedings of the 22nd International Conference on*

- World Wide Web*, 527-538. <https://doi.org/10.1145/2488388.2488435>
- Haroon, M., Wojcieszak, M., Chhabra, A., Liu, X., Mohapatra, P. and Shafiq, Z. (2023). [Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations.](#) *Proceedings of the National Academy of Sciences*, 120(50): e2213020120. <https://doi.org/10.1073/pnas.2213020120>
- Hosseinmardi, H., Ghasemian, A., Rivera-Lanas, M., Horta Ribeiro, M., West, R. and Watts, D.J. (2024). [Causally estimating the effect of YouTube's recommender system using counterfactual bots.](#) *Proceedings of the National Academy of Sciences*, 121(8): e2313377121. <https://doi.org/10.1073/pnas.2313377121>
- Hussein, E., Juneja, P. and Mitra, T. (2020). [Measuring misinformation in video search platforms: An audit study on YouTube.](#) *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1): Article 48. <https://doi.org/10.1145/3392854>
- Krippendorff, K. (2018). [Content Analysis: An Introduction to Its Methodology](#), 4th edn. Sage, Thousand Oaks, CA.
- Metaxa, D., Park, J.S., Robertson, R.E., Karahalios, K., Wilson, C., Hancock, J. and Sandvig, C. (2021). [Auditing algorithms: Understanding algorithmic systems from the outside in.](#) *Foundations and Trends in Human-Computer Interaction*, 14(4): 272-344. <https://doi.org/10.1561/11000000083>
- Nosek, B.A., Alter, G., Banks, G.C., et al. (2015). [Promoting an open research culture.](#) *Science*, 348(6242): 1422-1425. <https://doi.org/10.1126/science.aab2374>
- Sandvig, C., Hamilton, K., Karahalios, K. and Langbort, C. (2014). [Auditing algorithms: Research methods for detecting discrimination on internet platforms.](#) Paper presented at Data and Discrimination: Converting Critical Concerns into Productive Inquiry, preconference of the 64th Annual Meeting of the International Communication Association, Seattle, WA.
- Tomlein, M., Pecher, B., Simko, J., Srba, I., Moro, R., Stefancova, E., Kompan, M., Hrckova, A., Podrouzek, J. and Bielikova, M. (2021). [An audit of misinformation filter bubbles on YouTube: Bubble bursting and recent behavior changes.](#) *Proceedings of the 15th ACM Conference on Recommender Systems*, 1-11. <https://doi.org/10.1145/3460231.3474241>
- Vombatkere, K., Mousavi, S., Zannettou, S., Roesner, F. and Gummadi, K.P. (2024). [TikTok and the art of personalization: Investigating exploration and exploitation on social media feeds.](#) *Proceedings of the ACM Web Conference 2024*.