

Manufactured Grassroots

Lessons for Civil Society from a Reported
AI Influence Attempt Ahead of Kenya's
2027 Elections



“

The public needs institutions capable of establishing what happened, distinguishing evidence from inference, and pursuing accountability where responsibility can be substantiated.

About Orodata

Orodata Science for over 13 years, has combined data science, technology, design, and research to improve institutional reforms, foster media innovation, bolster health systems, and facilitate access to inclusive and rich data.

We use design to simplify information to empower people, institutions, and organizations at different levels with knowledge that leads to the development of solutions that impact the lives of millions.

Country Operations: Aitokhuehi John

Creative Development: Justin Chuka

Research: Ugochukwu Dominic

Editor: Uche Oti

Contact: partnerships@orodata.org +234 8034333640, +234 7038645898

Address: 42 Montgomery Rd, Sabo Yaba, Lagos 101245

Disclaimer: This brief draws on Anthropic’s reported findings concerning an AI-enabled influence attempt in Kenya. It does not independently verify the underlying activity, its sponsorship, public reach, or effects. References to a “pro-government” operator describe reported political alignment and should not be interpreted as evidence of government direction, funding, or authorisation.

The discussion of potential democratic harms is analytical and does not establish that those harms occurred in this case. Policy recommendations are the brief’s own proposals, not positions attributed to Anthropic. They address deceptive misrepresentation of public participation, not legitimate political advocacy, anonymous expression, coordinated civic organising, or AI use in itself.

Executive Summary



Executive Summary

Anthropic's September 2026 threat intelligence report highlights pro-government actors in Kenya using AI to generate fake grassroots social media content ahead of the 2027 general election. This disclosure underscores a threat deeper than mere misinformation: **the deployment of AI to disguise coordinated political campaigns as genuine public sentiment.**

For civil society, the central issue is beyond the accuracy of the individual posts. One must ask if the identities, independence, and apparent diversity of the voices behind them are being misrepresented.

The case requires careful interpretation. Anthropic's account is a company-reported finding, not an independently verified assessment by this brief. Preparing deceptive messages does not establish that they were published, reached authentic audiences, changed political opinions, or affected electoral prospects. Likewise, describing an operator as "pro-government" does not establish government direction, funding, or authorisation. The disclosure should be treated as a basis for investigation and preparedness, and not proof of successful political influence.

The potential democratic harm nevertheless warrants attention.

Centrally produced messages presented as unrelated citizens' views could create a misleading impression of public consensus, making a contested policy appear widely accepted or a particular grievance seem unrepresentative. Such signals could affect not only individual citizens but also journalists, public officials, and civic organisations interpreting public sentiment. These are plausible risks rather than demonstrated effects of the Kenyan attempt. They matter in everyday policy debates as well as during elections.

The case also highlights a gap between corporate detection and public accountability. AI providers may observe planning and content production within their services, while evidence of publication, distribution, and audience exposure may sit elsewhere. Anthropic acknowledges that understanding activity beyond its platform requires external evidence.

Effective scrutiny therefore depends on connecting corporate findings with platform data, open-source research, and the contextual knowledge of journalists and communities. Disruption by a company can be valuable without resolving public-interest questions about responsibility, reach, or the adequacy of its assessment.

This brief recommends shared civic investigation capabilities, rather than expecting every civil society organisation to establish its own technical laboratory. The priorities are to:

- **Provide accessible investigative support:** connect community organisations, journalists, technical researchers, and legal or policy specialists through clear referral and collaboration arrangements.
- **Make corporate disclosures usable:** seek clear accounts of observed conduct, the basis for attribution, the scope of visibility, and unresolved questions, alongside evidence that can safely support independent scrutiny.
- **Adopt common evidence standards:** preserve relevant material securely and distinguish preparation from publication, exposure from persuasion, and political alignment from sponsorship.
- **Protect legitimate participation:** focus on demonstrable deception and misrepresented coordination, rather than treating anonymity, shared messaging, political organising, or AI use as inherently suspicious.

The policy lesson is neither that every influence attempt succeeds nor that an attempt without demonstrated impact can be ignored. It is that civil society needs the sustained capacity to establish what happened, communicate uncertainty responsibly, and pursue accountability where the evidence supports it.

The objective extends beyond the removal of deceptive posts. It is to strengthen the public's ability to distinguish genuine participation from manufactured appearances of consent.



The objective extends beyond the removal of deceptive posts. It is to strengthen the public's ability to distinguish genuine participation from manufactured appearances of consent.

The Disclosure



The Disclosure

The disclosure describes an attempt to make organised political messaging appear to be independent citizen expression. It provides a basis for examining how AI could be used to support deceptive participation, but it should not be treated as proof of successful influence or government involvement.

The evidentiary distinction is important: **the report establishes that Anthropic has publicly described this activity; this brief does not independently verify the underlying operation.** The company's findings are a source for investigation. They are not a substitute for corroboration.

Reported preparation is not demonstrated public influence

Preparing messages, publishing them, reaching an audience, and changing people's views are distinct stages. Evidence of one does not automatically establish the next. Generated posts may never be published; published posts may receive little attention; and visible engagement does not necessarily indicate persuasion.

The reported preparation of fake grassroots posts therefore supports concern about the operator's intended method. It does not, by itself establish

the existence of a functioning network of accounts, substantial public exposure, or effects on political behaviour. Claims that the activity shaped Kenyan public opinion or affected electoral prospects would require additional evidence.

The same discipline should apply to assessments of failure. A company's conclusion that an operation achieved little or no meaningful reach should be attributed to that company and assessed against its stated evidence and visibility. An absence of demonstrated impact is not necessarily proof that no exposure occurred.

Political alignment is not evidence of government direction. The description "pro-government" concerns the political orientation of the reported messaging. It does not establish who commissioned, financed, authorised, or supervised the activity. An operator could support a government without acting on its instructions.

Attributing responsibility would require corroborating evidence linking the activity to particular individuals or institutions. Until such evidence is available, the brief should avoid describing the case as a government-run campaign. Equally, the absence of an established link should not be

Source: Anthropic's September 2026 threat intelligence report.

converted into a definitive claim that no such connection existed. **Unresolved sponsorship should remain an explicitly unresolved question.**

Corporate visibility is valuable but incomplete

Anthropic acknowledges this boundary in its discussion of influence investigations, stating that its visibility ends once an operation is live and that understanding subsequent activity requires open-source research, cross-platform information, and public reporting. Its disclosures should therefore be read with attention to which conclusions rest on activity observed within its services and which depend on external evidence.

This does not infer that we dismiss corporate findings. It instead provides reason to distinguish their evidentiary basis and recognise that no single organisation necessarily holds the complete record. Establishing public distribution, audience exposure, or sponsorship may require information beyond an AI provider's own systems.

A disclosed case does not establish the scale of the wider problem

The Kenyan case cannot, on its own, show how common AI-enabled political



manipulation is in Kenya, how many actors are involved, or whether such attempts generally succeed. Nor does it demonstrate that AI-generated messages are more persuasive than messages produced without AI.

Anthropic presents its report as a selection of notable cases, rather than a representative account of all misuse. The disclosure can illustrate a method and identify questions requiring attention, but it cannot support population-wide conclusions about the prevalence or effects of that method. Broader claims would require a wider and more systematic evidence base.

The appropriate conclusion is a bounded one

The disclosure supports a focused concern: according to Anthropic, an operator sought to use AI to prepare political messages that misrepresented grassroots participation. It leaves separate questions about publication, distribution, audience response, and responsibility to be established through further evidence.

Civil society need not wait for proof of successful persuasion before taking an attempted deception seriously. But the response should remain proportionate to what is known.



Civil society need not wait for proof of successful persuasion before taking an attempted deception seriously. But the response should remain proportionate to what is known.



The Practical Relevance

The Practical Relevance

Fact-checking asks whether a claim is true. The reported Kenyan case raises another question: **is the public conversation what it appears to be?**

A political message need not contain a false statement to be deceptive. It may express a defensible opinion or accurately describe a policy while misleading readers about who is speaking. If centrally produced messages are presented as the spontaneous views of unrelated citizens, the deception concerns the origin and independence of those voices. Verifying the statements alone would not establish whether the participation behind them is genuine.

Manufacturing the appearance of consensus

People interpret public debate partly through the apparent number and diversity of those expressing a view. Messages that seem to come from different citizens can suggest that a position enjoys broader, more independent support than it actually does. The potential harm is therefore not limited to persuading someone to believe a particular claim. It also includes misleading people about what others believe.

An apparent groundswell of support

could make a contested policy seem widely accepted. Manufactured opposition could make a legitimate proposal appear broadly rejected. Either could affect perceptions of whether a grievance is shared, whether public disagreement is possible, or whether speaking out would attract support. These are plausible risks, not demonstrated effects of the reported Kenyan attempt. Preparing deceptive material does not establish that it circulated, reached authentic audiences, or changed their behavior.

Distorting the signals institutions receive

The relevant audience extends beyond individual voters. Journalists, public officials, and civil society organizations may encounter online reactions when assessing public priorities, identifying stories, or interpreting responses to government decisions. Where manufactured messages are mistaken for independent testimony, they could distort those assessments.

The reported electricity-tariff messaging illustrates why the issue matters beyond elections. Debates about public services and household costs are also settings in which apparent citizen support or opposition may carry political weight. Online

visibility is not a representative measure of public opinion, but it can still be treated as a signal of public sentiment. Concealed coordination risks making that signal misleading.

Protecting genuine grievances from dismissal

Deceptive participation may draw on real concerns to appear credible. Identifying fabricated accounts or centrally produced messages would not establish that the underlying grievance is invented, or that everyone expressing a similar view belongs to the same operation.

This distinction is essential. An investigation into manipulated debate about electricity costs, for example, should not become a reason to dismiss citizens' genuine experiences of affordability or service provision. Nor should uncertainty about some accounts cast suspicion over an entire political constituency. Scrutiny must separate evidence about deceptive conduct from judgements about the legitimacy of the issue being discussed

Distinguishing deception from legitimate organizing

Authentic participation does not require every speaker to use their real name,

compose entirely original messages, or act independently of others. Citizens may legitimately organize campaigns, share talking points, use pseudonyms, and rely on AI for drafting or translation. None of these practices, on its own, establishes manipulation.

The relevant distinction is between people coordinating to express their views and an operation misrepresenting centrally controlled voices as independent participants. Protecting that distinction is especially important where anonymity enables people to speak without fear of retaliation. A response intended to defend public debate should not impose identity-disclosure requirements or treat political agreement as evidence of wrongdoing.

The policy concern is therefore broader than the presence of "fake posts". It is whether the apparent composition of public debate can be trusted. **What is at stake is beyond citizens encountering accurate information. Can they and the institutions representing them judge whose voices they are hearing, how independently those voices are speaking, and what weight to give them?**

What It Means For CSOs



What It Means For CSOs

Public-policy monitoring matters alongside election observation

The reported electricity-tariff messaging makes the case relevant beyond electoral campaigning. Civil society organisations working on public services, consumer rights, and government accountability should consider how manufactured participation could distort the debates they monitor.

An apparent wave of support for a policy may warrant investigation where there is evidence of misrepresented identities or concealed coordination. It should not, however, be treated as suspicious simply because it supports the government or contradicts an organisation's position. What is worth considering is whether participation is being misrepresented, and not whether the views expressed are politically convenient.

Community knowledge is an investigative resource

Technical analysis alone cannot establish whether an account's claimed identity, local references, or description of community concerns is credible. Community organisations and local journalists can help researchers interpret language, assess contextual

inconsistencies, and understand whether online narratives reflect experiences on the ground. These contributions should inform investigation rather than be treated as conclusive proof: unfamiliar language or an unexpected political opinion does not establish that a speaker is fake. Effective scrutiny combines contextual knowledge with corroborating evidence

Shared capabilities should come with a clear division of responsibilities

Not every organisation needs to investigate account networks or trace digital infrastructure. Frontline civic groups could document concerns and preserve relevant public material; specialist researchers could examine coordination and distribution; journalists could seek corroboration and responses; and legal or policy organisations could assess appropriate accountability routes.

A shared model should make these roles accessible through clear referral arrangements, while establishing responsibility for evidence security, quality assurance, and decisions about publication. The objective is to make specialist support available without turning every CSO into a technical laboratory.

Public communication must distinguish suspicion from findings

Civil society's credibility depends on communicating uncertainty as carefully as it communicates risk. Reports should distinguish material prepared from material published, publication from audience exposure, and exposure from persuasion. They should also separate evidence of political alignment from evidence of sponsorship or direction. Where conclusions rely on a technology company's disclosure, that dependence should be explicit. Publishing methods, confidence levels, and unresolved questions would help prevent an investigation into manipulation from inadvertently spreading an unsupported allegation.

Accountability should extend beyond account removal

A platform takedown or an AI-provider suspension may interrupt activity without answering the questions that matter to affected communities. Who commissioned the messaging? Where was it distributed? Were genuine organisations or individuals impersonated? What evidence remains available for independent review? Civil society should pursue these questions where the evidence permits, while recognising that some may remain unresolved. Success should be assessed through the quality of verified findings, corrections, and institutional responses, not simply the number of posts flagged or accounts removed.



Frontline civic groups could document concerns and preserve relevant public material; specialist researchers could examine coordination and distribution; journalists could seek corroboration.

Policy Priorities



Policy Priorities

A. Establish shared civic investigation capabilities

Civil society organisations, newsrooms, election observers, and researchers should develop a shared investigative arrangement rather than expecting every organisation to maintain its own technical laboratory.

Its core functions should include:

- Receiving and assessing reports of suspected deceptive activity.
- Preserving publicly accessible evidence with dates, sources, and collection notes.
- Analysing potential coordination across accounts and platforms.
- Providing local-language interpretation, legal review, and digital security support.
- Publishing findings with clear confidence levels and documented limitations.

The arrangement should have independent governance, transparent funding, and consistent criteria regardless of the political interests involved.

B. Seek usable disclosures from AI developers and social platforms

Civil society should request information that enables verification, subject to privacy and security safeguards:

- What behaviour was directly observed, and during what period?
- What evidence supports descriptions such as “pro-government” or “coordinated”?
- Was the activity limited to preparation, or was public distribution established?
- What action did the company take, and what did that action demonstrably interrupt?
- Which relevant records can be preserved or shared with independent researchers?
- What remains unknown?

Restricting an operator’s access to one AI service should not be presented as proof that a wider political operation has ended.

C. Adopt common evidence and publication standards

Investigations should distinguish between suspected activity, corroborated coordination, attributed sponsorship, and demonstrated impact. Each requires different evidence.

AI-generated language, political agreement, anonymity, or high posting frequency should not independently justify an accusation. Public reports should explain alternative interpretations, offer an opportunity to respond where appropriate, and provide a correction mechanism.

Claims about impact require particular care: posting volume is not audience reach, and audience reach is not persuasion.

D. Protect civic freedoms within the response

Policy should target deceptive conduct, not political viewpoints or AI-assisted expression as such.

Civil society should resist blanket restrictions on anonymous accounts, broad prohibitions on AI-generated political speech, or measures that allow authorities to label dissent as manipulation without evidence. Research practices should minimise personal-data collection and protect

sources and ordinary users.

E. Fund sustained investigation, not only training

Funders should support the recurring costs of investigation: personnel, lawful data access, secure evidence storage, local-language expertise, legal assistance, and publication.

Success should be assessed through the quality and verifiability of investigations, institutional responses, and safeguards against false accusations, not simply the number of accounts flagged or removed.



Funders should support the recurring costs of investigation: personnel, lawful data access, secure evidence storage, local-language expertise, legal assistance, and publication.



Immediate Agenda For Kenya

Immediate Agenda For Kenya

The immediate priority is to turn the disclosure into a focused, independently assessable inquiry, not to presume that a successful campaign occurred. Kenyan civil society organisations, journalists, and researchers could pursue a shared investigation that separates what Anthropic reports from what can be corroborated publicly and what remains inaccessible. The following actions would provide a practical starting point.

1. Establish a common record of the reported findings

The first step should be a concise evidence record drawn directly from the original report. It should distinguish the operator's reported requests and content production from any evidence of publication, audience exposure, or sponsorship. Details circulating through screenshots, commentary, or media summaries should be checked against the primary source before being repeated.

The record should also identify the basis of each claim: whether it reflects activity Anthropic says it observed, a company assessment, independently corroborated evidence, or an unresolved question. This would help participating organisations work from the same factual baseline and avoid

gradually converting qualified findings into stronger allegations.

2. Request evidence that enables independent scrutiny

A coordinated request to Anthropic should seek clarification of the evidence supporting its account, including:

- The period covered by the observed activity.
- Whether the evidence concerns proposed messages, generated outputs, confirmed publication, or a combination of these.
- Whether any publicly accessible accounts or posts were linked to the activity, and on what basis.
- What evidence supports any assessment of audience reach or lack of impact.
- What information, if any, supports conclusions about the operator's affiliations.

Requests should prioritise information that can safely be shared, such as redacted examples, methodological explanations, and public identifiers where disclosure is justified. They should not assume that private user

records can or should be released. Where information cannot be shared, the resulting limits on independent verification should be made explicit.

3. Investigate whether material entered public circulation

Where usable examples or identifiers are available, specialist researchers could examine whether corresponding material appeared publicly and whether there is corroborating evidence connecting it to the reported operator.

Similar wording alone would not establish that connection. Shared slogans, repeated political arguments, and coordinated advocacy can arise legitimately. Investigators should assess multiple indicators together and document alternative explanations.

Any findings should distinguish publication from amplification and audience exposure. Public engagement figures may offer limited evidence of interaction, but they do not establish how many authentic people encountered the material or whether it persuaded them. If no matching material is found, that should be reported as a search result within defined limits, not proof that publication never occurred.

Policy should target deceptive conduct,

not political viewpoints or AI-assisted expression as such.

4. Bring Kenyan contextual expertise into the inquiry

Local journalists and civic organisations should help interpret the issues, language, and identities invoked in the messaging. For electricity-related narratives, this could include organisations working on consumer interests and public-service accountability; for election-related narratives, it could include electoral and political-rights specialists.

Their role would be to assess context and identify questions requiring corroboration, not to decide that an opinion is inauthentic because it appears unusual or politically inconvenient. If the inquiry identifies possible impersonation of a person or organisation, those affected should be approached through safe channels and given an opportunity to respond.

The investigation should also distinguish the misuse of a genuine public concern from the legitimacy of that concern. Evidence of manufactured participation must not become grounds for dismissing citizens' experiences.

5. Preserve evidence and establish safeguards before publication

Participating organisations should agree on procedures for preserving relevant public material, recording its source and collection date, and distinguishing original evidence from analysis. Access to sensitive information should be restricted according to investigative need.

Before naming an individual or institution, investigators should require corroboration proportionate to the allegation, assess foreseeable risks, and provide an appropriate opportunity to respond. Political alignment should not substitute for evidence of responsibility. Nor should an inquiry into concealed coordination expose unrelated anonymous speakers or publish personal information unnecessarily.

These safeguards would support both credible findings and the rights of people who might otherwise be wrongly associated with the activity.

6. Publish a bounded assessment and identify follow-up action

The immediate output should be a short public assessment that clearly separates:

- **Company-reported findings:** what Anthropic states it observed or concluded.
- **Independent findings:** what investigators could corroborate.
- **Unresolved questions:** what available evidence does not establish.
- **Investigative limits:** what could not be accessed, tested, or reliably inferred.

Further action should follow the evidence. Verified impersonation might warrant notification and remediation; substantiated deceptive account activity could support requests for platform investigation; and corroborated evidence of commissioning or financing could justify further accountability work. Where the evidence remains insufficient, the appropriate outcome may be to document the gap rather than make an attribution.

Conlusion



Conclusion

The reported Kenyan case is a warning about the manufacture of apparent public participation, not proof that AI successfully changed political opinion. Its significance lies in the attempted deception: centrally produced messages presented as independent grassroots expression. Anthropic's disclosure provides a starting point for scrutiny, but it should not be treated as an independently verified account of the activity's distribution, impact, or sponsorship. Pro-government messaging does not, by itself, establish government involvement.

For civil society, the challenge therefore extends beyond identifying false claims or detecting AI-generated text. It concerns the ability to investigate who is speaking, how activity is coordinated, and whether audiences are being misled about its origin. That work requires connecting evidence held by AI providers and social platforms with the contextual knowledge of journalists, researchers, and communities. Corporate disclosures can support this process, but they cannot substitute for independent public-interest investigation.

The practical response is to invest in shared civic investigation capabilities, rather than expect every organisation to build its own technical laboratory.

Accessible specialist support, secure evidence-preservation systems, common verification standards, and workable channels for requesting information would give civic organisations a stronger basis for investigating suspected manipulation and challenging unsupported assessments. Funders should support these capabilities as sustained infrastructure, not only as short-term interventions around elections.

These efforts must also protect the participation they seek to defend. Political agreement, coordinated advocacy, anonymous speech, and the use of AI are not inherently evidence of manipulation. Scrutiny should focus on demonstrable deception and concealed coordination, with clear safeguards against misidentification and the suppression of legitimate expression.

The central policy lesson extends beyond civil society proving every influence attempt succeeded before responding. It is that the public needs institutions capable of establishing what happened, distinguishing evidence from inference, and pursuing accountability where responsibility can be substantiated. **Simply fewer fake posts is not the goal here, but a stronger civic capacity to defend the authenticity of public participation.**

