BUILD UP

POLARIZATION FOOTPRINT METHOD

November 2025

This document outlines the method for a metric of affective polarization on social media platforms, which we are calling the “polarization footprint”.

The method was designed by Helena Puig Larrauri, Luke Thorburn, Caleb Gichuhi and Allan Cheboi, with inputs from Julie Hawke, Benjamin Cerigo, Daniel Burkhardt Cerigo, and Andrew Sutjahjo. The method benefited from early review and comments from Ravi Iyer, Jonathan Stray, Cathy Buerger, Smitha Milli, Medinat Malefakis, and Seid Muhiye Yimam. For questions, please contact team@howtobuildup.org

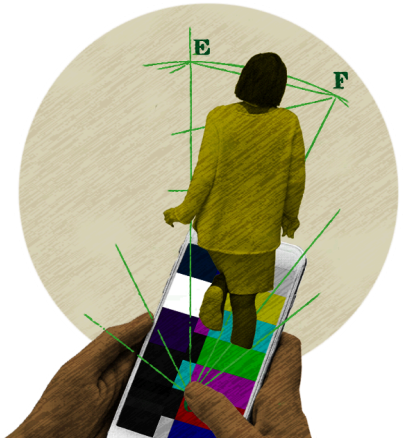


TABLE OF CONTENTS

1. OVERVIEW

The polarization footprint is a cross-platform measure that ranks social media platforms according to the prevalence of affective polarization present on the platform. The polarization footprint relies on observing content and relationships on social media platforms, and is designed in such a way as to be comparable across platforms. The methodology is explicitly designed to deliver defendable and confident measures of the minimum prevalence of polarization.

Showing platform users the numeric reflection and representation of polarizing behaviour may spur reflection and potentially policy or behavioural modifications. As such, the polarization footprint method:

- provides a template for how on-platform polarization can be measured using ecologically valid observational methods (that is, non-experimental, non-survey methods) by external researchers (i.e. researchers not working within platform companies);
- explores how league tables (comparing platforms) can incentivize platform reform and inform user choice; and
- spurs discussion of platform responsibilities with respect to conflict and polarization that informs policy about how the interaction of online platforms with societal conflict should be regulated, including by considering taxation on the polarization footprint.

To complement the polarization footprint, this method suggests running the Neely Social Media Index survey alongside the on-platform observational measure. The objective of the survey is to explore positive and negative user experiences on social media platforms. By running the survey together with the on-platform measure, we can derive some meaning from the relationship between their results, and understand the differences between perceived negative experiences and observable polarization.

This method was used by Build Up in Kenya in 2025. Where operational decisions were made specifically for the Kenyan context, this is indicated, in order to facilitate adaptation and replication to other contexts.

2. DEFINITIONS

2.1 Affective Polarization

Simply put, where issue-based polarization is where people disagree about issues, affective polarization is where people dislike and distance themselves from others because of their identity (including association with a position on an issue). Affective polarization is a dynamic process intertwined with conflict escalation, by which a self-reinforcing feedback loop separates ideologies or identity groups into increasingly distanced and aggregated adversaries. Affective polarization is relationship or identity-based in that its focus is the increasing dislike, distrust, and animosity towards those from other parties or groups. This

differs from issue-based polarization, which focuses on the disagreements or ideological distance between parties on policy areas.

Affective polarization is a key driver of conflict. Just as some conflict can in certain cases be constructive, issue-based polarization is not necessarily problematic. In contrast, affective polarization can increase the risk of escalation to violence by taking a conflict that is more specific and localized toward something more general, identity-based and antagonistic. Issue-based polarization becomes affective and intractable when we can't change what we think or say without losing core social networks or identities. Differing opinions do not necessarily lead to destructive conflict (the kind that erodes institutions and can turn violent), but the underlying practices that impact social cohesion often do. When affective polarization is wide-spread, conflict can become intractable: structures become rigid and de-escalation becomes very difficult.

The polarization footprint is a composite measure made up of three parts, each measuring a component of affective polarization – attitude polarization, norm polarization and interaction polarization – that are further defined below. Affective polarization is a human dynamic, and these three components seek to isolate one aspect of the dynamic so it can be measured separately. In reality, we know the three components interact, with causality likely running in many directions.

2.1.1 Attitude Polarization

Attitude polarization is characterized by perceptual shifts toward stereotypes, dehumanization, deindividuation, and vilification. Affective polarization by definition includes biased and negative attitudes about an out-group. These attitudes are often expressed as fixed and overgeneralized beliefs or notions about certain groups, identities, and their intersections. Referential examples include referring to an individual in plural or third person pronouns, speaking to an individual as one homogenous group with similar ideas, positions, or characteristics, or generalized blame and attribution for a context's shortcomings. Over time, a "rigidification" of an outgroup's identity can occur, making an integration of the two groups, and consequently the de-escalation of conflict or opposing narratives, more challenging to achieve.

We measure attitude polarization by examining whether the language used in social media posts denotes (i) negative stereotypes, (ii) dehumanization, (iii) deindividuation, (iv) vilification, or (v) calls to violence. The attitude polarization score is the percentage of posts and comments that contain this language. In effect, this means we are making a connection between attitude polarization and descriptive norms. Descriptive norms are defined as what we expect others to do, often as a result of what we see them do most often. We care about descriptive norms because they often impact the perceptual shifts that result in attitude polarization.

2.1.2 Norm Polarization

While attitudes are shaped by descriptive norms, and can determine what they believe is correct behavior, injunctive norms are the function of what we collectively understand. They represent a person's perception or idea of what behavior is socially acceptable or rewarded. Often, attitudes and norms are at play together, e.g. men's negative beliefs about women's roles also collectively shape social norms that reinforce the acceptance of behavior such as the ridicule and stereotyping of women in online discussions. Affective polarization is at play when combative norms of interaction reinforce the erosion of trust between people and towards representative social institutions. The erosion of trust is both an antecedent and a consequence of other polarization dynamics. When there are fewer interpersonal ties to counter negative stereotypes about the outgroup, and more in-group and institutional incentives for antagonism, people feel freer to employ more severe actions or rhetoric against the 'other'. When others see grievous actions or rhetoric, they acquire a basis of mistrust or negative expectations regarding the conduct of others. The extended impact of confirmed negative expectations changes the nature of groups and the self-protective ways they engage in discourse to reinforce competitive, defensive, apathetic, and combative norms for interaction.

We measure norm polarization by looking for challenges to polarized attitudes in the comments in a social media thread where either the post or one comment expresses attitude polarization. The norm polarization score is the percentage of threads where there is no challenge to polarized attitudes in the thread.

2.1.3 Interaction Polarization

As the structural middle falls out of a communication ecosystem, the lines of communication and everyday interaction that are normal to peaceful engagement are cut off. The reduction of conversation quality or quantity as a way to manage divergent viewpoints can signal a breakdown in meaningful engagement. Interpersonal relationships are deprioritized in relation to value or identity alignment, and networks are fragmented. Interaction polarization is the extent to which people are fragmented into dissimilar clusters, which impacts both the interests and affiliations of people, creating a self-reinforcing cycle of polarization.

Whereas we measure attitude and norm polarization based on individual behaviors (posting content, reacting to content), interaction polarization is a network-wide dynamic, not a post-level attribute. We measure interaction polarization as the degree of fragmentation in the hypergraph of post impressions and followed accounts – that is, how easy it is to predict based on one impression / follow what other impressions / follows a user will have.

2.2 User experiences on social media

The Neely Social Media Index survey explores positive and negative user experiences on social media platforms. Survey questions revolve around experience and usage of social platforms. Previous research (e.g., New Public's Civic Signals work) has found that learning new things and connecting with others are primary use cases for social platforms, so some survey questions measure those positive experiences. Negative experiences with social

platforms often revolve around content that is personally upsetting and content that is perceived to be bad for the world, so survey questions measure those experiences. Further background is available here. Survey questions were originally designed for the USA, and are adapted to account for contextually relevant topics and divisions.

3. DATA COLLECTION

3.1 Recruitment

We recruit participants to ensure **a representative sample of the population**.

- In Kenya, recruitment was conducted via ads placed on the five social media platforms: Facebook, Instagram, X, YouTube, and TikTok and the target sample size was 5000 complete responses (1000 per platform).
- For replication, recruitment could be conducted via a recruitment platform (Prolific, CloudResearch Connect, etc.) and the target sample size could be as low as 2500 complete responses (500 per platform).

Participants are **paid to compensate for their time** in completing a response, which takes approximately 20 minutes.

- In Kenya, participants were paid 1000 KSH
- For replication, payment will need to adapt to context

We **only target Android users or desktop browser users**. This is a pragmatic decision because we will be developing a custom app for data collection, and only have the resources to do this for one platform.

- In Kenya, the vast majority of people use Android[1], and we exclusively collected data via an Android app.
- For replication, we would assess the Android user base or consider a desktop browser option.

We use **demographic targeting (hard quotas)** to match demographics of the population who are online[2]. We match our sample to the *marginal* distributions of the target population along the following dimensions. The buckets we stratify along are given in square brackets. Options with an asterisk are monitored but not strict targets:

- age [18-34, 35-54, 55+]
- gender [female, male, non-binary*]
- province or equivalent level 2 administrative division [each province or equivalent][3]

[1] Stats from Feb 2024 suggest 87.87% Android, 2.75% iOS, 8.93% unknown.

[2] In Kenya, to estimate the demographics of the online population, we use data from the most recent Kenyan census (2019), appropriately weighting each province by the rates of internet use in each. We source this data from the Communications Authority of Kenya trends report. We believe this is the best reference available for the demographics of the target population. Similar sources will be required for replication.

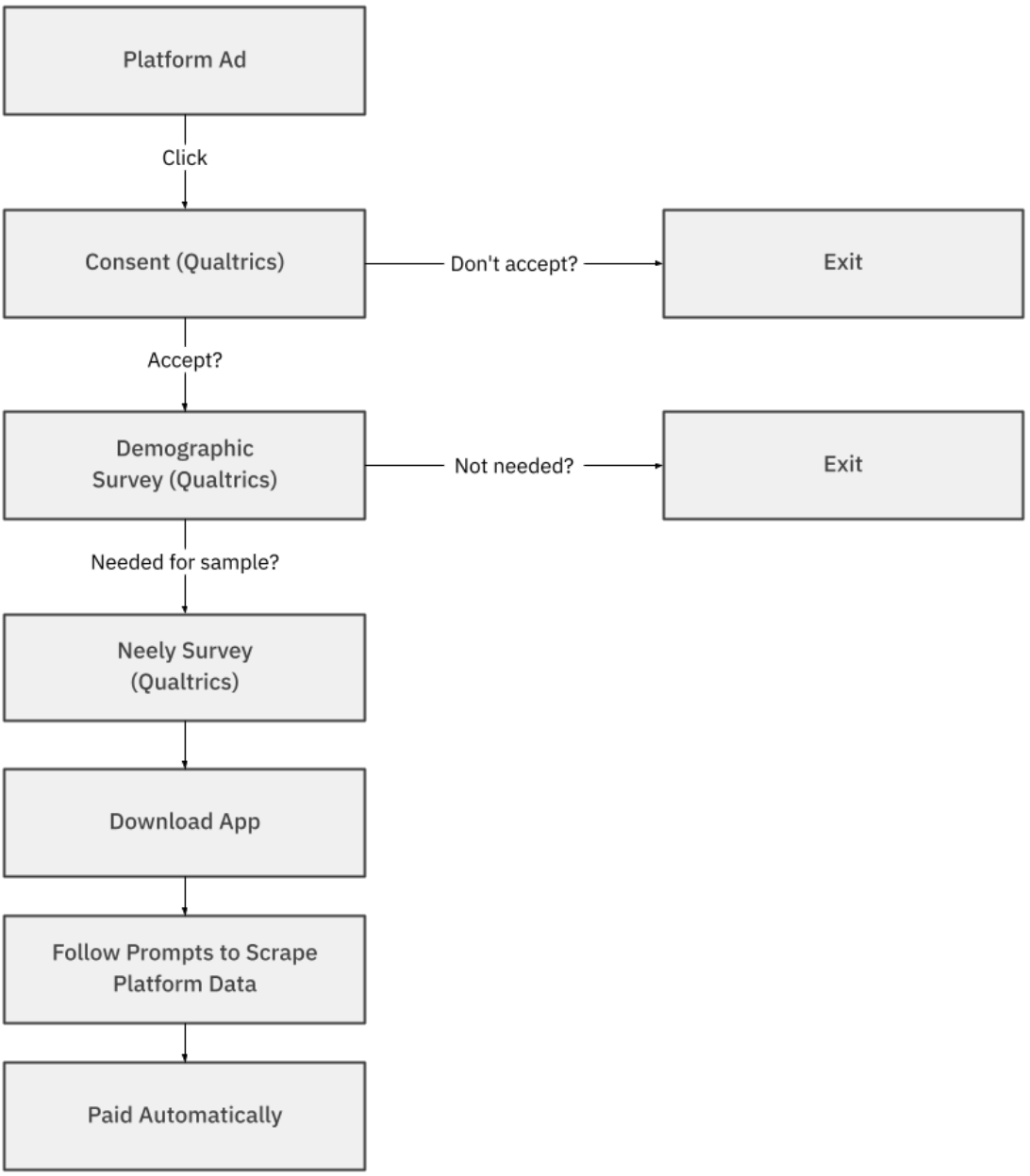
[3] The target distribution by administrative division is on the overall population distribution, weighted by % of internet use in each province.

We **monitor representativeness (soft quotas)** along the marginal distributions of the following dimensions, but they will not be hard targets for the representativeness of our sample.

- highest level of education completed
- ethnicity
- religion

3.2 Participant Flow

The flowchart below presents the overall flow for Kenyan participants once they have clicked on a recruitment ad. For replication, the flow would be similar, with the possibility of a different recruitment entry point, and web browser v. app data collection options.



3.4 Onboarding tasks

3.4.1 Recruitment content

Recruitment content (platform ad or other) covers:

- Study to understand social media in [country]
- Answer questions and share information about your social media use and earn [amount]
- You can answer on your phone, it will take about 20 minutes and you will receive the money instantly via [payment mechanism] after answering
- This study is run by Build Up [information]

3.4.2 Consent form

The consent form provides information on data privacy and protection, including:

- Repeats who runs the study and what its purpose is.
- Participation is voluntary, and you are free to stop at any time.
- What data are we collecting?
    - There are two parts to the study.
    - The first part is a survey, where we will ask you questions about your demographics, your experiences on social media. The demographics may include your age, gender, province, ethnicity, religion, and political attitudes. We will ask you for your [relevant payment details] so that we can pay you via [payment method], and delete your data if you ask us too.
    - In the second part, we will ask you to download an app and use it to log in to the social media platform where you clicked on the ad. The app will collect data about the first 100 posts that appear in your feed, including the contents of each post, the ID of the account that posted it, the first 100 comments on each post, and the IDs of the accounts that commented. Note that this may include posts from the people you have connected with on social media that are not visible publicly. We will also collect a list of the public accounts that you follow. Later, we will use this data to determine the 100 most followed public accounts among study participants. We will not collect any data from your direct messages or group chats.
- How will we protect data?
    - All the information will be protected by two-factor authentication, encrypted in transit, and only shared with the necessary researchers. The social media data collected by the app will be immediately and securely transferred to our secure server. It will not be stored on your device.
    - Your [payment data] will always be stored separately to the rest of the data, and will only be accessible by a designated researcher. In this way, your information will be anonymous during all our analysis.
- How long will we keep data for?
    - We will de-identify the survey data, and this non-identifiable data will be stored indefinitely in an academic data repository.

    - We will delete all social media data, and all sensitive or identifying information (i.e., your phone number and demographics) by the end of [data retention period].
- Who will have access to your data?
    - The information we collect will be managed by Build Up, and also accessed by [relevant researchers, if needed].
    - The de-identified survey data will be made available to other academics on request.
- What will we use your data for?
    - We will use your data to compare the patterns of behavior and kinds of content that are shown on different social media platforms. Anonymous, aggregate results will be reported publicly in academic papers and policy documents.
    - During our analysis, we will also use your social media data to train automated classifiers that can detect certain kinds of content. Build Up will make these classifiers publicly available under an open source license.
- You have 24 hours to complete the study.
    - The study will take about 20 minutes. If you do not complete it within 24 hours (from now), we will delete any data you have submitted and won't be able to pay you.
- You should delete the app once you have been paid.
    - By participating, you agree to delete the app once data collection has been completed and you have been paid. We will not maintain (and are not responsible for) the app after this point.
- You have a right to access or request removal of your information.
    - To withdraw from the study, or to access or request removal of your information, email [study email address].

Participants only proceed to the demographic survey if they click "Yes" to consent.

3.4.3 Demographic survey

The demographic survey starts by asking the hard quota questions. Participants only proceed if we detect that they are needed for our sample based on these quotas. The demographic survey then asks the remaining (soft quota) questions.

3.4.4 Neely Social Media Index Survey

Participants first complete (a contextualised version of) the Neely Social Media Index survey in Qualtrics. The Kenyan adaptation of the survey can be viewed in Annex 1. The survey includes attention checks.

3.4.5 Platform data collection

After completing the survey, participants are prompted to download a custom Android app, through which they are prompted to log in to the website of the platform on which they were recruited (if using platform ads) or the platform they indicated they use most regularly (if using

other recruitment method). The in-app browser will then scrape the relevant data that we need to collect for the study (documented in the first column of Table 1). In the background, the server will retrieve any supplementary data via platform APIs or secondary scrapes (documented in the second column of Table 1).

Overall, we scrape 500,000 posts (100,000 per platform), by scraping the first 100 posts of the private newsfeed of 5000 users (1000 per platform). We also scrape the first 100 comments for all posts collected (with the total number of comments dependent on how many comments we actually find per post).

Table 1 — Data collection pipeline for study. N (posts per user) = 100; K (comments per post) = 100. Only the data described in black text in each column was collected using that method. The data described in light grey text is only there for ease of comparison across columns.

	Data Collection Method	
Platform	**Scraped (during task)**	**Scraped (after via Apify or, for YouTube, via the official API)**

X	*For each of first N posts:*
	- ID/URL
	- text
	- account
	- timestamp
	- engagement counts
	- replies (first K):
	  - ID/URL
	  - text
	  - account
	  - timestamp
	  - engagement counts
	  - ID of parent post
	
	- public accounts followed
		For each of first N posts:
		- ID/URL
		- text
		- account
		- timestamp
		- engagement counts
		- replies (first K):
		  - ID/URL
		  - text
		  - account
		  - timestamp
		  - engagement counts
		  - ID of parent post
		
		- public accounts followed

Facebook	*For each of first N posts:*
	- ID/URL
	- text
	- account
	- timestamp
	- engagement counts
	- comments on non-public posts (first K):
	  - ID/URL
	  - text
	  - account
	  - timestamp
	  - engagement counts
	  - ID of parent post
	
	- public groups/pages followed
		For each of first N posts:
		- ID/URL
		- text
		- account
		- timestamp
		- engagement counts
		- comments on public posts (first K):
		  - ID/URL
		  - text
		  - account
		  - timestamp
		  - engagement counts
		  - ID of parent post
		
		- public groups/pages followed

	Data Collection Method	
Platform	**Scraped (during task)**	**Scraped (after via Apify or, for YouTube, via the official API)**

Instagram	*For each of first N posts:*
	- ID/URL
	- text
	- account
	- timestamp
	- engagement counts
	- comments on non-public posts (first K):
	  - ID/URL
	  - text
	  - account
	  - timestamp
	  - engagement counts
	  - ID of parent post
	
	- public accounts followed
		For each of first N posts:
		- ID/URL
		- text
		- account
		- timestamp
		- engagement counts
		- comments on public posts (first K):
		  - ID/URL
		  - text
		  - account
		  - timestamp
		  - engagement counts
		  - ID of parent post
		
		- public accounts followed

YouTube	*For each of first N posts:*
	- ID/URL
	- title
	- description
	- account
	- timestamp
	- engagement counts
	- comments:
	  - ID/URL
	  - text
	  - account
	  - timestamp
	  - engagement counts
	  - parent (reply/post)
	
	- public accounts followed (channels)
		For each of first N posts:
		- ID/URL
		- title
		- description
		- account
		- timestamp
		- engagement counts
		- comments:
		  - ID/URL
		  - text
		  - account
		  - timestamp
		  - engagement counts
		  - parent (reply/post)

TikTok	*For each of first N posts:*
	- ID/URL
	- text
	- account
	- timestamp
	- engagement counts
	- comments on non-public posts (first K):
	  - ID/URL
	  - text
	  - account
	  - timestamp
	  - engagement counts
		For each of first N posts:
		- ID/URL
		- text
		- account
		- timestamp
		- engagement counts
		- comments on public posts (first K):
		  - ID/URL
		  - text
		  - account
		  - timestamp
		  - engagement counts

	Data Collection Method	
Platform	**Scraped (during task)**	**Scraped (after via Apify or, for YouTube, via the official API)**
	- ID of parent post - public accounts followed	- ID of parent post - public accounts followed

3.4.6 Payment

Finally, participants are prompted to enter relevant payment details into a payment platform. Build Up will then reimburse participants (this will happen automatically and be near-instantaneous).

- In Kenya, participants were asked to enter their mobile number, for reimbursement via M-Pesa, a ubiquitous mobile payment method in Kenya.
- For replication, a similar mobile-enabled platform or the payment system used by any participant recruitment platform used (e.g., Prolific) would likely be most appropriate.

4. DATA ANALYSIS

The data analysis methodology is explicitly designed to deliver defendable and confident measures of the **minimum** prevalence of polarization across each of the three categories: attitude, norm and interaction.

4.1 Attitude Polarization

To measure attitude polarization, a team of trained experts use an agreed operational definition to annotate randomly selected posts and comments. Each post / comment is annotated by at least three experts, and the final labels (“Definitely Polarizing”, “Potentially Polarizing” “Not Polarizing”) are a conservative (tie-down) majority-vote of their annotations. At the platform level, the attitude polarization score is the percentage of posts and comments that are labelled as being definitely or potentially attitude polarizating.

We annotate a post or comment as containing attitude polarization by examining whether the language used denotes negative stereotypes, dehumanization, deindividuation, vilification, or a call to violence. We use text from the posts and comments collected during the study for annotation, concretely:

- Facebook: post text and comment text (incl. hashtags), if any
- Instagram: post description and comment text (incl. hashtags), if any
- X: post text and replies text (incl. hashtags), if any
- YouTube: video title and description and comment text (incl. hashtags), if any
- TikTok: video description and comment text (incl. hashtags), if any

As such, this means we have made the methodological decision not to classify videos on YouTube or TikTok, only classifying associated text, where it is available. This is justified, in part, by the experience of Build Up that attitude polarization is more likely to occur in comments than in posts, and that we report our prevalence estimates as lower bounds.

We use this annotated dataset to train a text classifier model, which is used to classify the entire dataset, a requirement for the norm polarization measure described in 4.2.

4.1.1 Operational definition

We define language as **definitely polarizing, potentially polarizing or not polarizing**, where "potentially" means:

- If this is repeated enough times, it would be polarising OR
- Some people could find this polarising

When applying this definition to classifying language in posts or comments, we are generally asking: **will this language change the reader's attitude towards another group?** Concretely:

1. **Stereotyping**: Does the text rely on oversimplified or fixed notions about a group or identity? This can include attributing positive characteristics to a group where:
    a. If there is a comparator and the opposite is negative, the text is definitely polarizing; e.g. x tribe is smarter than all other tribes
    b. If there is no comparator and the opposite is negative, the text is potentially polarizing; e.g. x tribe is smart
    c. If the opposite is value neutral, then it is not polarizing; e.g. x tribe are great dancers
2. **Dehumanization**: Are individuals or groups described in ways that strip them of human qualities (e.g., referring to people as animals, objects, or entities)?
3. **Deindividuation and Invalidation**: Does the text refer to an individual as if they represent an entire group or identity (e.g., using plural or collective terms like "they" or "those people")? Does the text invalidate people's identity's existence?
4. **Vilification**: Does the text assign generalized blame or attribute negative motives to a group for broader societal issues or shortcomings?
5. **Call to violence:** Does the text call for violence to be perpetrated against an individual or a group, including threats of violence?

The **negation of a group** is an identity marker. E.g. "not Kikuyu" is an identity marker.

Where a post or comment cannot be interpreted because there is **not enough context, then we label it as not polarizing**. We will not infer context.

We try to **avoid tone policing**, and as such we **only classify some insults as polarizing**, namely:

- If an insult is directed at an individual (not dehumanizing / blaming / about attributes), then it is not polarizing (it is an expression of anger); e.g. Fuck you, fuck off etc
- If an insult is directed at an individual and dehumanizing / blaming / about attributes, then it is potentially polarizing; e.g. gay, dog, a shit, monkey, prostitute, pig, a devil, vagina etc
- If an insult is directed at a group or attached to an identity, then it is definitely polarizing; e.g. all police officers are dogs, you're a police that makes you a dog

Posts that **describe / report polarization** using actual polarizing content are “potentially polarizing” if they repeat the actual polarizing language; they are “not polarizing” if they use neutral language.

Based on this agreed definition, we develop a **functional dataset** by searching social media platforms for examples of posts and comments that fit the different classes. We use this functional dataset to train expert annotators, and discuss any areas of uncertainty.

4.1.2 Development of an annotated representative dataset

We annotate a minimum of 9000 randomly selected examples[4], containing both posts and comments, using the following **labelling schema**.

text	The text of the post or comment.
message_id	A pseudo-identifier (anonymised) that can be used to link the data to other datasets for analysis.
attitude_polarization	The is the text is polarizing, with categories: ● 0 - Not Polarizing ● 1 - Potentially Polarizing ● 2 - Definitely Polarizing
is_comment	Integer flag (1/0).
contains_negative_stereotypes	Integer flag (1/0) to indicate whether text contains negative stereotypes.
contains_vilification	Integer flag (1/0) to indicate whether text contains vilification.
contains_dehumanization	Integer flag (1/0) to indicate whether text contains dehumanization.
contains_deindividuation	Integer flag (1/0) to indicate whether the text contains deindividuation.
contains_calltoviolence	Integer flag (1/0) to indicate whether the text contains a call to violence.
annotator_id	The ID of the annotator.
annotation_timestamp	Timestamp of the annotations, used crucially for enabling reannotation, keeping on the most recent annotations for a message.
seconds_to_annotate	The number of seconds it took to label the example by the annotator. Used for assessing if an annotator thought about the label and making estimates on how many we could do.
mark_for_review	A special boolean that indicates the example does not fit in the current definition and should be assessed by the team.

[4] In Kenya, our final annotated dataset had 9,773 examples.

dataset_name	To indicate which dataset this annotation is part of.

Annotation rounds

The examples that are chosen to be annotated are a random selection from the full scraped dataset and contain both posts and comments.

The annotation process is done in rounds to allow for the assessment of inter-rater reliability (see below) after each round and to follow the progress of the annotation in an iterative way.

- A round's sample size will start at 25 examples and increase by doubling in size for each iteration to a max size 1000.
- A round's dataset is created and added to our self-hosted Potato platform (online annotation system). It will be ordered differently for each annotator.
- Each example will be annotated for:
    - Attitude_polarization
    - Flags
    - Mark_for_review: this is a special label that lets annotators indicate that this example needs a review/further discussion, for instance if the example does not fit into the current description of attitude polarization.

The annotators are assigned a dataset of a particular size, and they each log in to the annotation platform online through their browser. Each annotator is served a post or comment, to which they add a labels (attitude_polarization and flags).

To leverage annotator time most effectively, and expecting that the majority (+90%) of posts are not polarizing, we use the following annotator strategy which significantly increases the number of posts/comments annotated in total while still maintaining highest confidence in the end labels for definitely and potentially polarizing posts/comments:

1. Do an annotation round where each post/comment is randomly assigned to three (a reduced number than the total) annotators - for a team of 8 total annotators and a round of 1000 posts/comments per annotator this gets through a total of 2666 posts/comments in that round.
2. Filter that round's resultant annotations to find posts/comments which were annotated as definitely or potentially *by any single annotator*. These posts/comments are then re-annotated by *all* annotators.

This strategy enables us to label approximately 3x more posts/comments, while still maintaining the highest confidence in the end label for all definitely and potentially polarizing posts/comments.

Once annotators have finished a complete round of annotations, inter-rater reliability (IRR) metrics are computed. Annotators then meet to discuss the examples with high levels of disagreement, or those that have been marked for review by an annotator. The result of this is continued moving towards more consistency and agreement in labelling between annotators.

The output of each review session is a list of *message_id*s which will be included again in the next annotation round, to be re-annotated (corrected).

Annotator (dis)agreement or inter-rater reliability (IRR)

We have two distinct scenarios that involve assessing annotator (dis)agreement. Firstly, during annotation, we want to identify posts/comments that created the largest disagreement between the annotators, so we can show these to the group of annotators for discussion and any amending of their annotations and possibly re-annotate them. Second, we need to report the level of agreement between the annotators as this determines how reliable the final (majority-vote) labels are for each post/comment.

One of the key problems is identifying why an annotation has a level of disagreement. It could be one of these reasons:

- Attitude polarisation does not have shared patterns and can't be consistently labelled
- Each annotator has sensitivities that might also be in a population
- The annotators are not in agreement about what the patterns that mean something is polarising
- There are mistakes in the annotation

We aim to analyse the annotations, assess what the reasons might be for the disagreement and take actions such as:

- Evaluate the definitions
- Leave disagreements as a sensitivity that is in the population that a model could consider
- Discuss as a group and clarify patterns for annotators
- Re-annotate examples

Identifying individual text items with largest disagreement

The purpose of *per (text) item disagreement* metrics are to enable the researchers and annotators to identify text items that have a high level of disagreement between the annotators labels, and thus surface those items for review, for further instruction/training, and for re-annotating (correcting) those items.

The exact specifications for the disagreement metrics are the code that computes them, found here.

The following are the first-tier disagreement metrics, and the highest scoring messages should be looked at for every annotation round:

1. *disagreement_polarization_distance_weighted:* conceptually, this is a measure how much the 0, 1, 2 attitude polarization classes differ between the annotators labels, where 0 <-> 2 is a bigger difference than 0 <-> 1 and 1 <-> 2. The higher the disagreement score, the higher the disagreement.

2. *disagreement_flags_avg*: conceptually, this is the average of how much the annotators disagree across all the flags. The higher the disagreement score, the higher the disagreement.

The following are second-tier disagreement metrics, which can be optionally used to identify less obvious issues in the annotation process:

1. *disagreement_{flag}*: For each of the flag/binary labels, e.g. “contains_vilification”. Enables finding disagreement on specific flags.
2. *disagreement_polarization_not_vs_definitely*: Disagreement but only between “Not Polarizing” and “Definitely Polarizing”, ignore “Potentially Polarizing” labels. Enables finding disagreements between “Not” and “Definitely” even for cases where the overall disagreement on polarization is low.
3. *disagreement_polarization_potentially_vs_definitely*: As per 2. But only between “Potentially Polarizing” and “Definitely Polarizing”
4. *disagreement_polarization*: As per *disagreement_polarization_distance_weighted* but without the distance weighting.

Practically, a spreadsheet with the examples, the labels given by each annotator, and the various disagreement scores, will be generated for the group to review.

Tracking and investigating inter-rater reliability (aggregated across all the text items) between annotation rounds

Between each round of annotation, we want to track and investigate annotator (dis)agreement, with the intention of identifying areas of this disagreement, and tracking that average agreement is improving. We will compute and assess the following:

1. Krippendorff’s α (ordinal distance weighting) for *attitude_polarization*
2. Gwet’s AC1 for each flag/binary label
3. Mean across all items for each of the per-item disagreement scores
4. Histogram of all items for each of the per-item disagreement scores

4.1.3 Calculation of results

The attitude polarization prevalence per platform is then calculated as follows:

1. Combine the annotation datasets from all annotation rounds conducted.
2. Filter to only retain the latest annotation by each annotator per post/comment. This removes the duplicate annotations that occur when the same posts/comments are re-annotated.
3. Apply the following label decision-rule:
    a. For each post/comment, count the number of annotators that gave it annotation definitely/potentially/not polarizing - these can be considered "votes".
    b. For each post/comment, give it the final label of whichever option has the most "votes" (most annotators annotated is as so).
    c. If there's a tie between two options, choose the final label to be whichever option is more conservative, i.e. definitely polarizing -> potentially polarizing ->

not polarizing, and definitely polarizing -> not polarizing. This is another element in ensuring the result is a defensible lower bound.

4. Compute the number of posts/comments that are thus definitely, potentially, not polarizing, per platform.
5. For attitude polarization prevalence we combine (sum) the number of definitely and potentially polarizing posts/comments, and then divide this number by the total posts/comments labelled, for each platform, to give a prevalence percentage. To compute the confidence intervals for the prevalence we use the counts with a 95% confidence standard Wilson score.

4.1.4 Development of a text classifier model

The text classifier training process is designed to enhance reliability, minimise bias and P-hacking potential, and support flexible experimentation within a reproducible methodology.

The labelled (post application of majority-vote tie-down decision rule) dataset is split into 30% training data, 20% validation data, and 50% test data. Models are trained on the training dataset, and then model optimisation decisions are based on the models' performance on the validation dataset (e.g. hyper parameter tuning, model architecture decisions, etc). Results on the test set are only computed once a best model has already been finalised based on performance on the validation dataset. The model's performance on the test data is the final reported model performance, representative of how well that model will perform on unseen data "in the wild". The models explored were fine-tuning pre-trained BERT models.

The experimental model that performs best is used to infer the attitude polarization of all posts and comments that have not been manually annotated through this process. This classified dataset is used, in addition to the human-annotated posts, to identify threads containing attitude polarization that will then be annotated for whether that attitude polarizing language is challenged, as described in Section 4.2. The absence of such challenges we call 'norm polarization.'

4.2 Norm Polarization

To measure norm polarization, a team of trained experts use an agreed operational definition to annotate a random sample of relevant post-comment threads. Only threads where either a post or a comment is labeled as "attitude polarization" are relevant to the norm polarization label. Each thread is annotated by two experts, and the final annotations are a conservative (tie-down) average of their annotations. At the platform level, the norm polarization score is the percentage of threads that are annotated as containing norm polarization.

We annotate a thread as containing norm polarization where at least one post / comment contains attitude polarization and there is no comment in the thread that challenges the post / comment expressing attitude polarization. We define a post-comment thread as follows:

- Facebook: a post and all the comments under it

- Instagram: a post description and all the comments under it
- X: a tweet and all replies under it
- YouTube: a video title and description and all the comments under it
- TikTok: a video description and all the comments under it

4.2.1 Operational definition

We define a challenge as any text that:

- Directly responds to the attitude polarization language (whether or not referencing the original post / comment), not necessarily be expressing the opposite opinion or viewpoint, but rather by taking a different stance on "othering":
    - Calling in (i.e. point to the polarizing language while using non-violent communication) always counts as a challenge
    - Calling out also counts as a challenge, as long as it does not include attitude polarization (i.e. addresses the individual and not the group)
- De-escalations / calling for a lowering of the tone are also considered challenges to attitude polarization.

Challenges cannot include attitude polarizing language.

4.2.2 Development of a randomly selected annotated dataset

To identify a random sample, we first apply the attitude polarization classifier to the entire dataset of scraped posts and comments, and create a sub-dataset that contains all the threads in which at least one post or comment is classified as probably_polarizing or definitely_polarizing.

Based on the size of this dataset, we conduct a power analysis to determine how many threads we need to label in order for our estimate of norm polarization to be a statistically significant estimate of norm polarization in the entire dataset[5]. We then randomly sample a larger number than this for annotation from the sub-dataset; we sample a larger number because we know some threads will be incorrectly classified by the model, and therefore will be discarded by annotators.

Two context experts then annotate every thread in sample using the following **labelling schema**:

- incorrect_attitude_polarization_label: used for instances where model has incorrectly identified attitude polarization. These threads are then filtered out in the norm polarization calculations.
- has_challenge: these are threads where there is a challenge to the present attitude polarization, and thus **are not** norm polarizing
- no_challenge: threads without any challenge to the present attitude polarization, thus norm polarizing

[5] In Kenya, this was 200 threads per platform.

The experts start by each annotating the same 100 threads. They then meet to discuss differences in their annotation, before proceeding to annotate the entire dataset.

4.2.3 Calculation of results

The final label of each thread is computed using the same decision-rule as applied to attitude polarization: majority-voting, with selecting the more conservative option in cases of tied votes (norm polarizing -> not norm polarizing). The prevalence and confidence intervals for the prevalence are computed in the same way as done for attitude polarization also.

At the platform level, the norm polarization score is the percentage of threads labeled as no_challenge.

4.5 Interaction Polarization

We construct a hypergraph, with one hyperedge for each account from which a post was seen in a feed. Participants share hyperedges for each account that they both saw (if any). We then discard all but the 100 largest hyperedges on each platform, corresponding to the 100 accounts recommended to the largest number of participants. If there are ties, we break them randomly to ensure exactly 100 accounts on each platform.

We then quantify interaction polarization as the total correlation between the indicator random variables that correspond to whether a randomly chosen participant belonged to a given hyperedge. Intuitively, this metric captures the degree of fragmentation in the sources of information people are recommended on each platform.

4.6 Polarization Footprint

We rank platforms for each type of polarization, and then use a Borda count to combine these rankings into a single overall ranking that can be used for a league table. Implicitly, this means we are treating each type of polarization as equally important.

For example, if platform A was observed to have the 2nd lowest levels of attitude polarization, the 3rd lowest levels of norm polarization, and the lowest degree of interaction polarization, its Borda score would be 6 (2 + 3 + 1). By ranking the platforms by this Borda score, we produce an overall ranking of the platforms by their level of affective polarisation.

4.7 Confidence & Robustness

Confidence intervals for the three primary metrics are computed using the following methods:

- For the prevalence of *attitude polarization*, and the prevalence of *norm polarization* (in attitude-polarizing threads), we use the Wilson score interval for estimating binomial proportions.

- For *interaction polarization*, quantified as total correlation (see Section 4.5), we use BCa bootstrap confidence intervals.

We conduct several **robustness checks** of the above metrics, observing how the results (and subsequent platform league table) varies with, e.g., different ways of resolving disagreement among human annotators, different ways of constructing the graph on which our measure of interaction polarization is based, and computing the primary metrics for different demographic subgroups.

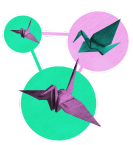
4.8 Survey & Combined Analysis

We report the results of all the survey questions, and break these down by relevant demographic groups. We also compare self-reported on platform experiences with observed content. Where the polarization footprint captures empirical 'ground truth', including exposure to polarizing content, as measured objectively by expert annotators, the survey captures self-reports of user experiences with polarizing content. As such, we conduct some high level analysis combining the results to look at whether empirical prevalence of polarizing content (polarization footprint) correlates with self-reported user experiences on platforms (survey).

5. RESEARCH ETHICS + INTEGRITY

In addition to the data protection & privacy guidelines described earlier in this method, we are guided by the following:

- **Review** — Before commencing any data collection, the project will undergo research ethics review at an accredited academic or research institution in the country.
- **Pilot** — Before committing to a final methodology, we conduct a pilot (a scaled down version of the full study with fewer participants) and make any required changes.
- **Pre-registration** — After the pilot and before commencing data collection for the main study, the study design and data analysis methods are pre-registered on OSF.
- **Transparency** – All results are made publicly available in a report and on a website.
- **Open source** – The code for producing the annotation servers, IRR metrics, prevalence metrics and classifier model are available here; the attitude polarization model is open source and made available on Hugging Face.



ANNEX 1: Neely Social Media Index survey (Kenya)

In the past 4 weeks, which of the following online services have you used? Check the box next to all that apply.
Katika wiki 4 zilizopita, ni huduma gani kati ya zifuatazo za mtandaoni umetumia? Weka alama kwenye kisanduku karibu na yote ambayo umetumia.

Options include: 1. Facebook 2. Twitter / X 3. Instagram 4. TikTok 5. Snapchat 6. YouTube 7. Reddit 8. WhatsApp 9. Email 10. LinkedIn 11. Pinterest 12. Dating Apps 13. Facetime 14. Text Messaging 15. Online Gaming 16. Twitch 17. Discord 18. Threads 19. Some other communications service 20. None of these services
Chaguo ni: 1. Facebook 2. Twitter / X 3. Instagram 4. TikTok 5. Snapchat 6. YouTube 7. Reddit 8. WhatsApp 9. Email 10. LinkedIn 11. Pinterest 12. Programu ya uchumba 13. Facetime 14. SMS 15. Michezo ya Mtandaoni 16. Twitch 17. Discord 18. Threads 19. Huduma zingine za mawasiliano 20. Hakuna kati ya huduma hizi.

For each service used ask:

How often have you used [*service*] in the past 4 weeks?

- ☐ Multiple times per day
- ☐ About once a day
- ☐ A few times per week
- ☐ About once a week
- ☐ Less than once a week
- ☐ I did not use [*service*] in the past 4 weeks

Je, umetumia [huduma] mara ngapi katika wiki 4 zilizopita?

- ☐ Mara kadhaa kwa siku
- ☐ Takriban mara moja kwa siku
- ☐ Mara chache kwa wiki
- ☐ Takriban mara moja kwa wiki
- ☐ Chini ya mara moja kwa wiki
- ☐ Sikutumia [huduma] katika wiki 4 zilizopita

In the past 4 weeks, have you personally witnessed or experienced something that affected you negatively on [*service*]?

- ☐ Yes
- ☐ No

Katika wiki 4 zilizopita, je, wewe binafsi umeshuhudia au kupatana na jambo ambalo lilikuathiri vibaya kwenye [huduma]?

- ☐ Ndio
- ☐ Hapana

If YES What was the impact of your negative experience(s) with [*service*] Check the box next to all that apply.

- ☐ It made me less likely to express myself online
- ☐ It negatively impacted my psychological well-being
- ☐ It reduced my trust in other people
- ☐ It reduced my trust in societal institutions
- ☐ It made me angry
- ☐ It worried me
- ☐ I felt unsafe
- ☐ I felt attacked
- ☐ It did not affect me a lot
- ☐ It annoyed me
- ☐ Other, please specify: [*text field]*

Kama umechagua NDIO, Je, matokeo ya kushuhudia au kupatana na jambo hilo lilio kuathiri vibaya katika [huduma] yalikuwa gani? Weka alama kwenye kisanduku kilicho karibu na yote yaliyo kutokea.

- ☐ Ilinifanya nipunguze uwezekano wa kujieleza mtandaoni
- ☐ Iliathiri vibaya ustawi wangu wa kisaikolojia
- ☐ It reduced my trust in other people
- ☐ Ilipunguza imani yangu kwa watu wengine
- ☐ Ilinikasirisha
- ☐ Ilinitia wasiwasi
- ☐ Nilihisi siko salama
- ☐ Nilihisi kama nimeshambuliwa
- ☐ Haikuniathiri sana
- ☐ Iliniudhi
- ☐ Nyingine, tafadhali taja: [nafasi ya maandishi]

If YES Did your experience(s) on [*service*] relate to any of these topics?
Check the box next to all that apply.

- ☐ Medical/health information
- ☐ Politics
- ☐ Crime
- ☐ Local news
- ☐ Personal finance
- ☐ Religion
- ☐ Climate / environmental issues
- ☐ Entertainment
- ☐ None of the above

Kama umechagua NDIO Je, hii matokeo yako kwenye [huduma] ili husiana na mojawapo ya mada hizi? Weka alama kwenye kisanduku karibu na yote yanayo husiana.

- ☐ Taarifa za matibabu/afya
- ☐ Siasa
- ☐ Uhalifu
- ☐ Habari za mitaa
- ☐ Fedha za kibinafsi
- ☐ Dini
- ☐ Masuala ya hali ya hewa / mazingira
- ☐ Burudani
- ☐ Hakuna kati ya zilizo hapo juu

If YES In a sentence or two, please describe one experience on [*service*] that personally affected you negatively. Please do not include any names of people, or your location, or other identifying information in what you write.
Kama umechagua NDIYO, Katika sentensi moja au mbili, tafadhali eleza tukio moja kwenye [huduma] ambalo lilikuathiri vibaya kibinafsi. Tafadhali usiweke majina yoyote ya watu, au eneo lako, au maelezo mengine ya utambuzi katika unachoandika.

In the past 4 weeks, have you witnessed or experienced content that you would consider bad for the world on [*service*]? (examples could include content that is misleading, hateful, or unnecessarily divisive)?

- ☐ Yes
- ☐ No

Katika wiki 4 zilizopita, je, umeshuhudia au kupitia habari ambazo unafikiri kuwa mabaya kwa ulimwengu kwenye [huduma]? (mifano inaweza kuwa habari inayopotosha, ya chuki, au yenye inagawanya watu)?

- ☐ Ndio
- ☐ Hapana

If YES What negative impact do you feel your experience(s) with [*service*] could have on the world? Check the box next to all that apply.

- ☐ It could increase political polarization
- ☐ It could increase hate, fear, and/or anger between groups of people
- ☐ It could increase the risk of violence
- ☐ It could misinform or mislead people
- ☐ It likely would not have much of an effect
- ☐ Other, please specify: [*text field]*

Kama umechagua NDIO Je, ni athari gani mbaya unahisi matokeo yako kwa [huduma] inaweza kuleta duniani? Weka alama kwenye kisanduku karibu na yote yanayowezekana kwa maoni yako.

- ☐ Inaweza kuongeza mgawanyiko wa kisiasa

- ☐ Inaweza kuongeza chuki, hofu, na/au hasira kati ya vikundi vya watu
- ☐ Inaweza kuongeza hatari ya vurugu
- ☐ Inaweza kupotosha au kupotosha watu
- ☐ Inawezekana haitakuwa na athari nyingi
- ☐ Nyingine, tafadhali taja:[nafasi ya maandishi]

If YES Did your experience(s) on [*services*] relate to any of these topics? Check the box next to all that apply.

- ☐ Medical/health information
- ☐ Politics
- ☐ Crime
- ☐ Local news
- ☐ Personal finance
- ☐ Religion
- ☐ Climate / environmental issues
- ☐ Entertainment
- ☐ None of the above

Kama umechagua NDIO Je, matokeo hii ya habari hizi kwenye [huduma] yalihusiana na mojawapo ya mada hizi? Weka alama kwenye kisanduku karibu na yote yanayo husiana..

- ☐ Taarifa za matibabu/afya
- ☐ Siasa
- ☐ Uhalifu
- ☐ Habari za mitaa
- ☐ Fedha za kibinafsi
- ☐ Dini
- ☐ Masuala ya hali ya hewa / mazingira
- ☐ Burudani
- ☐ Hakuna kati ya zilizo hapo juu

If YES In a sentence or two, please describe one experience on [*services*] with content that you would consider bad for the world. Please do not include any names of people, or your location, or other identifying information in what you write.

Ikiwa umechagua NDIO, Katika sentensi moja au mbili, tafadhali eleza tukio moja kwenye [huduma] kuhusu habari ziliemo, ambalo unaona kuwa mabaya kwa ulimwengu. Tafadhali usiweke majina yoyote ya watu, au eneo lako, au maelezo mengine yanayo weza kutumika kutambulisha mtu katika unachoandika.

In the past 4 weeks, have you experienced a meaningful connection with others on [*services*]? (examples could include exchanging emotional support or bonding over shared experiences)

- ☐ Yes
- ☐ No

Katika wiki 4 zilizopita, je, umepitia au kushuhudia muunganisho wa maana na wengine kwenye [huduma]? (mifano inaweza kuwa kubadilishana msaada wa kihisia au uhusiano kutokana na shughuli mbalimbali ulioshirikishwa)

- ☐ Ndio
- ☐ Hapana

> If YES In a sentence or two, please describe one experience on [*services*] where you meaningfully connected with others. Please include who you connected with. Please do not include any names of people, or your location, or other identifying information in what you write.
> Kama umechagua NDIO, Katika sentensi moja au mbili, tafadhali eleza tukio moja kwenye [huduma] ambapo uliungana na wengine kwa njia nzuri. Tafadhali elezea uliyeunganishwa naye, kwa mfano mzazi au rafiki au mwalimu na kadhalika. Tafadhali usiweke majina yoyote ya watu, au eneo lako, au maelezo mengine yanayo weza kutumika kutambulisha mtu katika unachoandika

In the past 4 weeks, have you learned something that was useful or that helped you understand something important on [*services*]?
Katika wiki 4 zilizopita, je, umejifunza kitu ambacho kilikuwa muhimu au kilichokusaidia kuelewa jambo muhimu kwenye [huduma]?

> If YES In a sentence or two, please describe one experience on [*services*] where you learned something useful or which helped you understand something important. Please include what you learned. Please do not include any names of people, or your location, or other identifying information in what you write.
> Kama umechagua NDIO, Katika sentensi moja au mbili, tafadhali eleza tukio moja kwenye [huduma] ambapo ulijifunza jambo muhimu au lililokusaidia kuelewa jambo muhimu. Tafadhali elezea ulichojifunza. Tafadhali usiweke majina yoyote ya watu, au eneo lako, au maelezo mengine yanayo weza kutumika kutambulisha mtu katika unachoandika

In the past 4 weeks, have you used applications that use AI to generate human-like text or code, such as ChatGPT, Llama, or Gemini?
Katika wiki 4 zilizopita, je, umetumia programu zinazotumia Akili bandia kutengeneza maandishi au msimbo unaofanana na binadamu, kama vile ChatGPT, Llama, au Gemini?

- ☐ Ndio
- ☐ Hapana

If YES You said earlier that you had used applications that use AI to generate human-like text or code, such as ChatGPT, Bard, or Bing Chat. What did you use them for? Please select all that apply.

- ☐ Out of curiosity
- ☐ For entertainment
- ☐ For social connection
- ☐ To learn something new about the world
- ☐ For tasks at work
- ☐ For school-related tasks
- ☐ To generate additional income (other than your regular work)
- ☐ To gather information or explore details about a specific health condition or treatment
- ☐ To create content for social media
- ☐ To assist in personal tasks, such as planning activities, trips, getting ideas for gifts, etc.
- ☐ To improve communications (for instance, help in writing emails, letters, etc.)
- ☐ As a tool for mental health, such as working through thoughts or emotions
- ☐ To help with creative pursuits, like writing stories, scripts, music, etc.
- ☐ Other, please specify: [*text field]*

Ikiwa umechagua NDIO Ulisema hapo awali kwamba ulikuwa umetumia programu zinazotumia Akili Bandia kutengeneza maandishi au msimbo unaofanana na binadamu, kama vile ChatGPT, Bard, au Bing Chat. Ulizitumia kwa ajili gani? Tafadhali chagua zote zinazotumika.

- ☐ Kutokana na udadisi
- ☐ Kwa burudani
- ☐ Kwa uhusiano wa kijamii
- ☐ Ili kujifunza kitu kipya kuhusu ulimwengu
- ☐ Kwa majukumu ya kazi
- ☐ Kwa kazi zinazohusiana na shule
- ☐ Kuzalisha mapato ya ziada (kando na kazi yako ya kawaida)
- ☐ Kukusanya taarifa au kuchunguza maelezo kuhusu hali mahususi ya afya au matibabu
- ☐ Ili kuunda habari kwa mitandao ya kijamii
- ☐ Kusaidia katika kazi za kibinafsi, kama vile kupanga shughuli, safari, kupata mawazo ya zawadi, na kadhalika.
- ☐ Ili kuboresha mawasiliano (kwa mfano, usaidizi wa kuandika barua pepe, barua, na kadhalika.)
- ☐ Kama chombo cha afya ya akili, kama vile kupumzisha na kuongoza mawazo au hisia
- ☐ Ili kusaidia kwa shughuli za ubunifu, kama vile kuandika hadithi, michezo ya kuigiza, muziki, na kadhalika.
- ☐ Nyingine, tafadhali taja: [nafasi ya maandishi]

If YES Please rate how useful or not useful your use of applications that use AI to generate human-like text or code was to you.

a. Not at all useful
b. Not very useful
c. Somewhat useful
d. Very useful
e. Extremely useful

Ikiwa umechagua NDIO, Tafadhali eleza kama utumiaji wako wa programu zinazotumia Akili Bandia ulikuwa muhimu au usio na manufaa kwako kutengeneza maandishi au msimbo unaofanana na binadamu.

a. haina manufaa hata kidogo
b. Haina manufaa sana
c. Ina manufaa kwa kiasi fulani
d. Ina manufaa sana
e. ina manufaa kwa hali ya juu

If YES Please rate how harmful or not harmful your use of applications that use AI to generate human-like text or code was to you.

a. Not at all harmful
b. Not very harmful
c. Somewhat harmful
d. Very harmful
e. Extremely harmful

Ikiwa umechagua NDIO Tafadhali eleza jinsi utumiaji wako wa programu zinazotumia Akili Bandia kukuletea maandishi au msimbo unaofanana na binadamu ulivyokuwa unadhuru au usiodhuru.

a. Bila madhara hata kidogo
b. Bila madhara sana
c. Inadhuru kwa kiasi fulani
d. Inadhuru sana
e. Inadhuru vibaya sana

Artificial intelligence computer programs are designed to learn tasks that humans typically do. How concerned or not concerned are you about the increased use of artificial intelligence computer programs in daily life?

Programu za kompyuta za akili za Bandia zimeundwa ili kujifunza kazi ambazo wanadamu hufanya kwa kawaida. Je, ni kiasi gani unajali au huna wasiwasi kuhusu ongezeko la matumizi ya programu za kompyuta za akili bandia katika maisha ya kila siku?

Options include: 1. Very concerned 2. Somewhat concerned 3. Not very concerned 4. Not at all concerned 5. No opinion

Chaguo ni: 1. Ninajali sana 2. Ninajali kwa kiasi fulani 3. Sijali sana 4. Sijali hata kidogo 5. Sina maoni

Artificial intelligence computer programs are designed to learn tasks that humans typically do. How excited or not excited are you about the increased use of artificial intelligence computer programs in daily life?
Programu za kompyuta za Akili za Bandia zimeundwa ili kujifunza kazi ambazo wanadamu hufanya kwa kawaida. Je, ni kiasi gani umesisimka au huna msisimko kuhusu ongezeko la matumizi ya programu za kompyuta za akili bandia katika maisha ya kila siku?

Options include: 1. Very excited 2. Somewhat excited 3. Not very excited 4. Not at all excited 5. No opinion
Chaguo ni: 1. Nimesisimka sana 2. Nimesisimka kwa kiasi fulani 3. Sina msisimko sana 4. Sijasisimka hata kidogo 5. Sina maoni

What are you most excited about regarding how AI could benefit your work and/or humanity? (text)
Je, ni nini kinacho kufurahisha sana kuhusu jinsi Akili za Bandia inaweza kufaidi kazi yako na/au ubinadamu? (maandishi)

What are your biggest concerns about how AI could negatively impact your work and/or humanity? (text)
Je,ni nini kinacho kutia wasiwasi mkubwa kuhusu jinsi Akili Bandia inaweza kuathiri vibaya kazi yako na/au ubinadamu? (maandishi)

Polarization Dependent Variables
Vigezo tegemezi vya migawanyiko ya jamii

Please indicate how you feel towards people who support the same political party that you support.

10 - you feel very favorably or warm toward them
9
8
7
6
5 - Neutral
4
3
2
1
0 - you feel very unfavorable or cold

Tafadhali onyesha jinsi unavyohisi kuhusu watu wanaounga mkono chama cha kisiasa unachokiunga mkono pia.

10 - Unahisi vizuri kuwahusu au upendeleo kwao
9
8
7

6
5 - hauegemei upande wowote kuwahusu
4
3
2
1
0 - Unahisi vibaya sana au huwajali

Please indicate how you feel towards people who support a different political party other than the one you support.

10 - you feel very favorably or warm toward them
9
8
7
6
5 - Neutral
4
3
2
1
0 - you feel very unfavorable or cold

Tafadhali onyesha jinsi unavyohisi kuhusu watu wanaounga mkono chama tofauti cha kisiasa na kile wewe unachokiunga mkono.

10 - Unahisi vizuri kuwahusu au upendeleo kwao
9
8
7
6
5 - hauegemei upande wowote kuwahusu
4
3
2
1
0 - Unahisi vibaya sana au huwajali

Please indicate how you feel towards people who are from the same religious group as your own.

10 - you feel very favorably or warm toward them
9
8
7
6
5 - Neutral
4

3
2
1
0 - you feel very unfavorable or cold

Tafadhali onyesha jinsi unavyohisi kuhusu watu ambao ni wa kundi moja la kidini kama lako.

10 - Unahisi vizuri kuwahusu au upendeleo kwao
9
8
7
6
5 - hauegemei upande wowote kuwahusu
4
3
2
1
0 - Unahisi vibaya sana au huwajali

Please indicate how you feel towards people who are from a different religious group to your own.

10 - you feel very favorably or warm toward them
9
8
7
6
5 - Neutral
4
3
2
1
0 - you feel very unfavorable or cold

Tafadhali onyesha jinsi unavyohisi kuhusu watu ambao wanatoka katika kundi la kidini tofauti na lako.

10 - Unahisi vizuri kuwahusu au upendeleo kwao
9
8
7
6
5 - hauegemei upande wowote kuwahusu
4
3
2
1
0 - Unahisi vibaya sana au huwajali

Please indicate how you feel towards people who are from the same ethnic group as your own.

10 - you feel very favorably or warm toward them
9
8
7
6
5 - Neutral
4
3
2
1
0 - you feel very unfavorable or cold

Tafadhali onyesha jinsi unavyohisi kuhusu watu wanaotoka kabila moja na lako.

10 - Unahisi vizuri kuwahusu au upendeleo kwao
9
8
7
6
5 - hauegemei upande wowote kuwahusu
4
3
2
1
0 - Unahisi vibaya sana au huwajali

Please indicate how you feel towards people who are from a different ethnic group to your own.

10 - you feel very favorably or warm toward them
9
8
7
6
5 - Neutral
4
3
2
1
0 - you feel very unfavorable or cold

Tafadhali onyesha jinsi unavyohisi kuhusu watu wanaotoka kabila tofauti na lako.

10 - Unahisi vizuri kuwahusu au upendeleo kwao
9
8
7
6

5 - hauegemei upande wowote kuwahusu
4
3
2
1
0 - Unahisi vibaya sana au huwajali