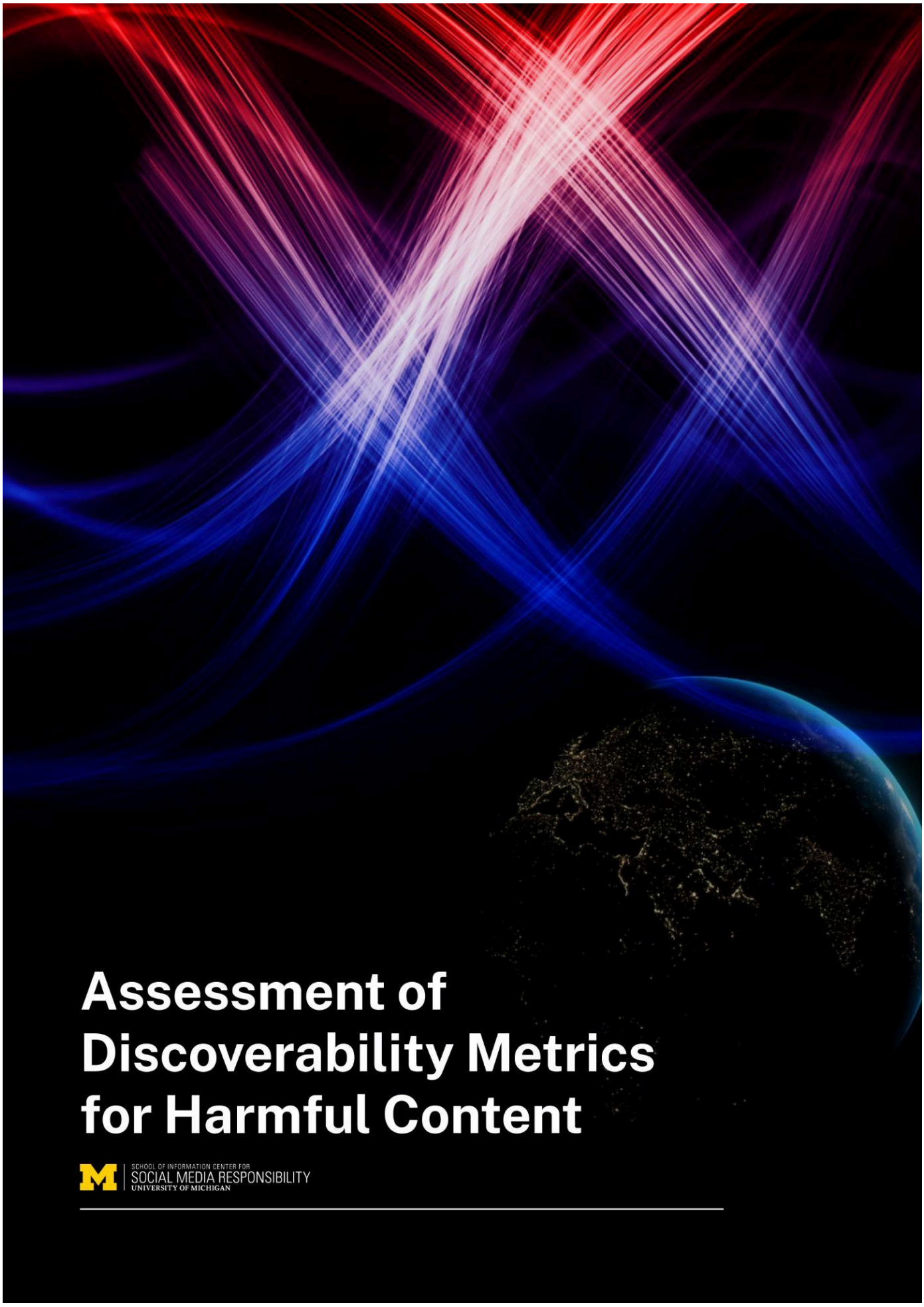

Assessment of Discoverability Metrics for Harmful Content

SCHOOL OF INFORMATION CENTER FOR
SOCIAL MEDIA RESPONSIBILITY
UNIVERSITY OF MICHIGAN

Assessment of Discoverability Metrics for Harmful Content

Paul Resnick, Siqi Wu, and James Park
University of Michigan Center for Social Media Responsibility
December 22, 2023

Executive Summary

Many stakeholders are interested in tracking prevalence metrics of harmful content on social media platforms. TrustLab tracks several metrics and has produced a report for the European Commission's Code of Practice. One of TrustLab's prevalence metrics, which they refer to as "discoverability," is calculated by simulating a set of user searches, classifying the results as harmful or not, and reporting the proportion of harmful results.

At TrustLab's request,[1] the University of Michigan Center for Social Media Responsibility (CSMR) has identified key concerns and considerations in producing and interpreting any discoverability metric, and some possible approaches for addressing these concerns. There is a family of possible discoverability metrics, each based on alternative design choices. This report analyzes the impacts of design alternatives on two key principles from measurement theory [1], validity and reliability, alongside a third principle, "robustness to strategic actors."

- Validity: the accuracy of the metric – whether it correctly measures the thing that it is supposed to measure. For discoverability metrics in particular, a key element of validity is comparability – the extent to which comparisons across platforms, countries, and time periods are meaningful. For example, is harmful content more prevalent on X or YouTube? Is it more prevalent in Slovakia or Poland? Did the prevalence decline on YouTube in Poland since last quarter?
- Reliability: the consistency of the metric – a measure could be accurate on average but have a high degree of variability between repeated individual measurements. In that case, any particular measurement would have to be treated as unreliable.
- Robustness to strategic actors: whether, for example, a platform could manipulate or game the discoverability metric without changing what real users experience on the platform.

[1] TrustLab provided funding for this project. The contract terms included a commitment that the University of Michigan was free to make its report public and provided an opportunity for TrustLab to demand the removal of confidential information only. TrustLab staff provided information about how it produces its discoverability metric. The CSMR authors drafted the report. TrustLab staff commented on drafts. The authors take full responsibility for the final contents.

The Family of Discoverability Metrics

TrustLab offers an intuitive definition of discoverability as “a measure of how easily a platform surfaces harmful content to a user searching for sensitive topics.” All processes for measuring discoverability implement a version of the same overarching process:

- Run some queries that simulate user searches for sensitive topics and harvest some search results.
- Label search results as positive or negative (i.e., harmful or non-harmful, depending on the project) via human labeling or an automated classifier.
- Compute the fraction of items labeled as positive.

Alternative versions of the discoverability metric will differ in the details of these steps, including: which search queries are generated; how they are generated; and how search results are labeled. They may also involve an adjustment to the way the fraction is calculated in order to take into account known patterns of errors in the harmfulness labels.

We note that there are also two alternative approaches that are out of scope for this report. Those alternatives would start by collecting either a set of personal feeds that users would experience on the platform or a random sample of all content from the platform (perhaps weighted by the view-count metric to focus on more popular items). This report is restricted to discoverability metrics calculated from the results returned by explicit user search queries.

Validity

TrustLab has operationalized the following plain-English definition as a precise numerical quantity: “the fraction of items that are harmful among the top-five search results on a selected set of sensitive topics,” as carried out by a set of human search agents and with harm judged by a human rater. This operationalization leads to several threats to validity. First, the searches may not reflect topics that real users would actually search for and may not allow for meaningful comparisons between countries, platforms, and time periods. Second, the simulated searches may not yield the same results that real users would encounter in practice because of platform personalization. Third, labeling errors may yield incorrect prevalence estimates of how harmful the results are. In what follows, we explore these threats to validity in detail as well as potential approaches for mitigating these threats.

Query Choices

The prevalence of harmful content resulting from a random selection of queries is likely to be quite low on most platforms, which would make it prohibitively expensive to get a precise estimate of prevalence. Thus, any practical discoverability metric is likely to focus on a subset of queries that might yield harmful content. We call these “sensitive searches” as a shorthand.

It would be ideal to select a random sample of the search queries that platform users actually ran on the platform, filtered based on whether the searches were sensitive or not. However, TrustLab or other external collectors of metrics have no access to the platform's search logs. Even an internally collected metric would face the challenge of identifying whether each search query was sensitive. Thus, in practice, a discoverability metric starts with the selection of some queries. The exact query set could have a large impact on the observed prevalence of harmful content in the results. We identify four design dimensions for the procedure for selecting search queries.

Harmful vs. sensitive topics. A key conceptual consideration is whether the selection process simulates an end-user deliberately seeking harmful content (i.e., searching on a harmful topic) or an end-user conducting a search where the results might include harmful content that the user was not specifically looking for (i.e., searching on a sensitive topic).

Selection of queries on a topic. Given a description of a topic, it would be possible to systematically select a few best queries on the topic. A potential semi-automated process could be: 1) generate candidate queries; 2) execute them; 3) label the results as on-topic or off-topic; 4) select the query with the most on-topic results (highest precision); 5) repeat steps 1-4 but only consider on-topic search results that are novel (not returned by previously included queries). As an alternative to systematic assembly of the best queries on a topic, human experts can simply select one or a few that are emblematic of the kinds of searches they think people might do on the topic.

Universal topics vs. custom topics for each country. A single set of topics could be selected for use in all countries. After selecting a query or queries for each topic, the set of topics would then be translated into all human languages used in each of the countries. At the other extreme, the process of selecting topics could be separate for each country, reflecting the diversity of topics of concern in different countries.

Frequency of changing topics. A different set of topics could be selected in each time period based on what the current interest is. Another approach is to select more general topics that will be of interest for a prolonged period of time and track those same topics for many weeks.

Effects of Query Choices on Metric Comparability

If a metric is valid, it should enable meaningful comparisons—between countries, platforms, and time periods. Using exactly the same set of queries in all countries, on all platforms, for all time, would ensure the highest degree of comparability. However, universal, unchanging queries result in some disadvantages. First, some topics are primarily of local interest, such as that country's current elections. Second, some platforms may have different search mechanisms (e.g., treating multi-word phrases as requiring exact matches vs. fuzzy matches vs. a boolean combination of single word matching). Finally, user interest in topics may wane over time.

Thus, instead of defining the discoverability metric in terms of results of a particular set of queries, it may be appropriate to define it in terms of a single, repeatable process that can be

applied in each platform, country, and time. Then, the metric would be properly interpreted as ***measuring the fraction of harmful content returned by searches on topics selected through that process***. So long as it makes sense to think of that process as producing conceptually the same thing when run on different platforms, countries, and time periods, the discoverability metrics can be compared.

For example, for a particular time period and country, one could use search data from a major search platform (e.g., Google search trends) to determine the most popular queries in a country. To restrict to only harmful topics, label the query strings and only use the harmful ones as input to the repeatable process. To restrict to sensitive topics, label the search results and only retain the ones with at least one harmful result in the first ten. The retained queries could then be clustered to indicate larger topics and a topic description could be generated for each cluster of queries. Each topic could then be turned back into a set of best queries for each social media platform, through the process described in the previous section.

This selection process would yield a discoverability metric that can be described as "measuring the fraction of harmful content returned by the best queries on the most popular sensitive topics in a country." It would be reasonable to compare that metric between countries, even if the most popular topics were different in the two countries. It would also be reasonable to make comparisons over time, even if different topics became popular. Finally, it would be reasonable to make comparisons between platforms, even if their search mechanisms are different, because the same process of selecting precise queries for the same topics was repeated on all platforms.

If, however, the process for selecting topics and converting topics to queries cannot be described in simple terms like "selecting the best queries on the most popular sensitive topics," it will be harder to consider the resulting measures comparable. In particular, if the selected topics can only be described as "*some* sensitive search topics" and the queries as "*some* queries on those topics," then comparing between countries or comparing between time periods will be problematic. It would be much better if the process could be described as selecting all search topics or the best search topics according to some criterion, even if that criterion is not overall popularity.

Topic-Specific Discoverability Metrics

A discoverability metric could be computed separately for each topic. The process of turning topics into specific high-precision queries could still be carried out separately for different platforms, ensuring comparability between platforms. Comparisons between countries or over time would be made one topic at a time, and only for topics that are present in both countries or time periods.

The advantage of this approach is that it does not require any way of describing two different sets of topics as comparable. One disadvantage is that it would require a large enough sample size of query results *on each topic* to yield a reasonably precise estimate of the proportion of harmful results. Another disadvantage is that it does not provide a way to make any aggregate

statement about the overall discoverability of harmful content, only about discoverability of harmful contents on some pre-selected topics.

Aggregating Topic-Specific Discoverability Metrics: An Index Approach

To provide a single aggregate discoverability metric, a basket of pre-selected topics could be used to define an index. The overall metric would, initially, be the average of the discoverability metric over all topics. The basket of topics would change over time, but slowly. Changes in the discoverability metric would then reflect increases or decreases in harmful search results for topics while they are in the basket, but would not reflect changes in the composition of topics.

Consider an analogy to stock market indexes, such as the Dow Jones Industrial Average (DJIA). The DJIA at any time is a basket of 30 publicly-traded companies; the set of companies changes, but only rarely. On day one, the index is computed as the average share price of the 30 companies. It goes up or down to reflect the change in portfolio value that an investor would have if they invested in a way that tracks the index (e.g., in an index fund).

Today, however, the DJIA is not simply the average price of the current 30 companies. Suppose a company with a $100 share price is removed from the basket and replaced by a company with a share price of $50. This replacement should not change the index (investors in an index fund would not lose or gain money as a result of this replacement). Intuitively, the only thing that should affect the DJIA is changes in the value of shares *while they are in the basket*. The DJIA accomplishes this by adjusting the denominator in a scenario like this, effectively making it as if there are fewer than 30 companies in the basket [2].

By analogy, an aggregate discoverability index would designate an initial set of topics. In each time period, the percentage of harmful results on each topic would be measured and the index would be calculated initially as the average of the harmful percentage over the topics.

If a new topic is added, however, it would be done in a way that does not change the index in the first time period when it is added. This can be achieved by adjusting the divisor, analogous to what the DJIA does.

A special note for this index is that it would be important to make changes to the basket based on the popularity of topics, not on the discoverability scores of topics. If a topic is added to the index because it becomes popular, the discoverability index could go up or down over time, depending on whether the prevalence of harmful results goes up or down after the topic is added to the index. If, instead, topics get added because they yield a high prevalence of harmful search results, then we would always add near a peak prevalence, and the index would continually go down.

Appendix A provides a sample scenario for calculation of the index and adjustment of the divisor.

Personalized Search Results

For a given set of search queries, it would be ideal to get the results from the user searches using those queries that naturally occurred on the platform. However, an external metric collector like TrustLab does not have access to the platform's full search logs.

Instead, TrustLab hires contractors to create accounts and run their queries. An alternative would be to have automated bots run queries, but TrustLab avoids that in order to comply with platforms' terms of service. TrustLab asks the contractors to create accounts to use only for searches on behalf of TrustLab. This could pose a threat to the validity of the metric, if the search results the contractors receive are very different from the results that regular platform users receive when running those same search queries.

Social media platforms have varying levels of personalization. They may produce different results based on geolocation, even within a country. A user's previous history of searching, browsing, liking and/or commenting behavior could be used to customize results. The same search could even yield different results at different times of day. Thus, TrustLab contractors' accounts, with their limited behavioral history that is not typical of other platform users and their different locations and timing, could be getting systematically different search results.

There is no way to entirely rule out this potential mismatch except to get access to actual search logs. However, robustness checks can be carried out.

One robustness check would be to assess the extent of personalization on a platform. For example, contractors (or bots) could simulate different patterns of use on different accounts, and then the same queries could be run on multiple accounts. If the results were not highly personalized, concerns about this threat to validity would be reduced.

Even if it turns out there is significant personalization, it might or might not have a big effect on the fraction of results that are harmful. TrustLab could have a variety of contractors, in different locations and with different lengths of account history (i.e., recent hires vs. seasoned), conduct searches. If the fraction of harmful content is consistent across contractors, that would increase confidence that their results may also match the results of organic searches on the platform.

Labeling Errors

TrustLab has written definitions of harm for each of three domains: misinformation, incitement to violence, and hate speech. Each creates a binary notion of harm for each single label: an item is either harmful or not. Even so, there are items where the application of the written definitions is not straightforward, and even trained human raters may disagree about whether an item is truly harmful. For example, imagine an item where if 100 trained human raters labeled it, 70 of them would label it as harmful.

These rater disagreements can pose a threat to validity. The threat to validity is slightly different depending on how we conceptualize the underlying ground truth of how harmful an item is. Below we describe the alternatives. Under an objective ground truth conceptualization, in order to make a correct prevalence estimation from human labels, a calibration process is needed. Under a subjective ground truth conceptualization, the calibration step is not needed, but the meaning of the discoverability metric changes from the fraction of results that are objectively harmful to the average subjective harmfulness of the results.

Objective Ground Truth Interpretation

In the objective ground truth interpretation, an item is either fundamentally harmful or not. In the example above, the item is truly harmful, but only 70% of trained evaluators are able to correctly detect that. With this interpretation, the discoverability metric is intended to capture the fraction of items that are truly harmful from all sensitive searches.

Unfortunately, with this interpretation it is not safe to use just a single human evaluation for each item. It would be safe if the number of items mislabeled as harmful equals the number of items mislabeled as non-harmful. However, the misclassifications may be asymmetric, and thus it is not safe to ignore them.

To see why, imagine that for truly harmful items on average 70% of raters label them as harmful, while for non-harmful items 90% of raters label them as non-harmful. This asymmetry leads to an undercount of harmful items. In particular, if there are half harmful and half non-harmful items, we can expect that out of all obtained human labels, only 40% of labels are harmful (.5 * 0.7 + .5 * 0.1 = .4). Thus, we would incorrectly get a measurement of 40% harmful items even though 50% of them truly were harmful.

One way to reduce this threat to validity is through calibration. If we think of a single human rater's label as an error-prone prediction about the objective ground truth, we can calibrate that prediction with a training sample. A large number of trained human raters label each item in the sample; the majority vote of the labels is a good proxy for the ground truth. It could still be incorrect, but with a large enough sample it will be unlikely, assuming carefully crafted annotation guidelines. The proxy ground truth can then be used to calibrate the meaning of a single rater's label. In the scenario above, with half harmful and half non-harmful items, seven out of every eight harmful labels come from a truly harmful item. Thus, if an item's only label says that it is harmful, there is an 87.5% chance that the item is truly harmful. Similarly, if an item's only label says that it is not harmful, there is a 25% chance that it is harmful.

Assume that we find such calibrated probabilities on a training sample, we obtain one human label per item, and we observe 40% of labels are harmful and 60% are non-harmful, as we would expect in this scenario. Instead of incorrectly reporting an estimate that 40% of items are harmful, using the calibrated probabilities would yield a correct estimate of the fraction of truly harmful items (.4 * .875 + .6*.25 = .5).

A similar calibration process can also be used for automated classifiers if TrustLab substitutes or supplements human labels with LLM-generated labels. The average of the classifier's calibrated outputs would yield an unbiased estimate of the fraction of harmful items. For details on using calibration to account for the error profile of black box classifiers, see [3]. Whether using just a single human label for each item or an automated classifier, the key is to calibrate the signal from either of them using a training sample that is labeled by a large number of raters to obtain the objective ground truth labels.

Subjective Ground Truth Interpretation

In the subjective ground truth interpretation, an item's harmfulness is a number between 0% and 100%. In our example, the item is 70% harmful because 70% of trained human raters think that the item's content satisfies the written definition. With subjective ground truth, the discoverability metric should be interpreted as the "percentage of harmful judgments" on the returned search items.

Here, the threat to validity from labeling errors is reduced. It is sufficient to get a single rater's label for each item, because the percentage of items whose single label is harmful is an unbiased estimator of the overall percentage of harmful labels if there were many labels per item [4]. There is also no need for a calibration step. In the example above (half items that are 70% harmful and half that are 10% harmful), the average subjective harmfulness is 40% (.5 * 0.7 + .5 * 0.1), which is a correct report of the expected percentage of harmful judgments that would be collected if there were many raters for many items.

If an automated classifier is used, its outputs can be calibrated against human raters' labels for a set of training items, but the calibration step still only requires one human label for each item in the training set. Then, the average of the classifier's calibrated outputs will again give an unbiased estimate of the average harmfulness of a set of items.

Reliability

Reliability indicates the consistency of the intended metric if the estimation process were repeated multiple times. We identify three factors that may impact reliability, and we describe procedures that an external metric collector like TrustLab could implement occasionally in order to estimate the reliability of the discoverability metric.

The selection of sensitive topics and queries. If the metric is defined in terms of a process for selecting topics rather than a predefined set of topics, running that process multiple times could yield different topics. Similarly, if a process is run to convert topics into specific queries, running that process multiple times on the same set of topics, even for the same country, platform, and time period, could yield different queries.

Collection of search results. Even if personalized search results are found not to pose a large threat to validity, they could introduce variability in measurements.

Human inter-rater agreement rate for harmful content. Even if ratings are calibrated and adjustments are made to prevalence estimates, as described in the Labeling Errors section, collecting just a single label for each result could introduce variability in measurements.

TrustLab could measure the (in)consistency of the empirical estimates by occasionally conducting repeated measurements. For the process of selecting topics or queries, this would require running the entire process multiple times, including not only selecting queries but also collecting search results and labeling the content for harm. For the inconsistency introduced by collecting search results, the same queries could be reused but run multiple times by different searchers, as described in the section on validity threats from personalized search results.

The inconsistency due to differences between human raters' labels could be estimated by collecting multiple human labels for a sample of items. Indeed, with several labels per item, a bootstrapping procedure could be used to generate a confidence interval around the measured fraction of harmful content. Each bootstrap sample would consist of a set of items (sampled with replacement) from the original dataset and a single label selected at random from the several labels available for each item. Each sample would yield an estimated fraction of harmful content, and the interval covering 95% of the results could be treated as a confidence interval, following the percentile bootstrap procedure [5].

Robustness to Strategic Actors

Suppose a platform wanted to reduce its discoverability metric without changing the extent to which it surfaces harmful content to users on its platform. Analogous to how Volkswagen programmed their diesel engines to activate emission controls only during emissions testing, how might a platform "game" the metrics [6]? The key would be to identify searches originating as part of the discoverability metric measurements and treat them differently than they treat the same searches conducted organically by users.

One way would be to identify the accounts of TrustLab's contractors who conduct the searches. (As noted previously, TrustLab does not rely on bots to conduct searches because platforms' Terms of Service often prohibit that.) Thus, TrustLab should ensure that their contractors are instructed not to reveal information that would identify them as TrustLab contractors. Even if they do not identify themselves, however, their behavior patterns may be identifiable. For example, it may be highly unusual for an account on a social media platform to primarily conduct searches and not engage in other typical behaviors, such as browsing, posting, liking, and commenting.

Even if a platform can identify when searches are being conducted as part of the discoverability metric measurement, it may not be easy for them to provide search results that are on topic but not harmful. If they could, they would presumably be doing so for searches by everyone, not just for those by TrustLab contractors. Thus, TrustLab may want to check whether the platform is returning results that are related to the particular search queries; if they are not, it could be an indication that the platform is providing specially sanitized results to game the metric.

Rather than identifying searchers, a platform might identify the set of specific queries that are run and treat those queries specially. If the set of queries is small, a platform might be able to optimize for those queries without reducing harmful content on other similar queries on the same topic. To check for this, it might be useful for TrustLab to include, in each time period, a few new queries on any topic that was continued from the previous time period. If the platform returns much less harmful content for the repeated queries than for the new queries on the same sensitive topics, it could be an indicator that the platform is optimizing for the specific queries they detected previously.

Similarly, if a platform knows the set of topics, it may optimize its performance for those topics. For comparability it is desirable to keep the same topics over time, and for transparency it will be desirable to reveal the topics, even if not the specific queries. Thus, it may not be possible to prevent gaming of this kind. The best solution here may be to include a comprehensive enough set of sensitive topics whereby platforms reducing the harmful results on such topics would constitute significant progress toward reducing harmful results for all sensitive topics.

References

[1] Allen, M. J., & Yen, W. M. (2001). Introduction to measurement theory. Waveland Press.
[2] Dow Divisor - Overview, History, and How to Calculate. https://corporatefinanceinstitute.com/resources/equities/dow-divisor/
[3] Wu, S., & Resnick, P. (2024). Calibrate-Extrapolate: Rethinking Prevalence Estimation with Black Box Classifiers. arXiv preprint arXiv:2401.09329.
[4] Resnick, P., Kong, Y., Schoenebeck, G., & Weninger, T. (2021). Survey equivalence: A procedure for measuring classifier accuracy against human labels. arXiv preprint arXiv:2106.01254.
[5] Diciccio, T. J., & Romano, J. P. (1988). A review of bootstrap confidence intervals. Journal of the Royal Statistical Society Series B: Statistical Methodology, 50(3), 338-354.
[6] Volkswagen emissions scandal - Wikipedia. https://en.wikipedia.org/wiki/Volkswagen_emissions_scandal

Appendix: Topic Index Sample Calculation

We use a concrete example with toy data to demonstrate the index calculation. At time *t1*, suppose that we have 9 topics. Express each topic's discoverability score as a decimal value (e.g., if 20 out of 100 search results were harmful, the score would be 20% or 0.2). The initial index is simply the average of those discoverability scores. For example, if the sum of topic discoverability scores at *t1* is 0.9, the index would be 0.9 / 9 = 10%. The initial divisor is thus 9, the number of topics.

At *t2*, suppose that we add a new topic, whose discoverability score is 0.2 (20%). Moreover, suppose that the discoverability scores for all the existing topics are unchanged. We would like changes in the index to reflect only changes in discoverability within topics, not changes in the composition of topics. If we simply calculate the average, we would obtain an index of (0.9 + 0.2) / 10 = 11% . Since the scores of all the stable topics were unchanged, we would like the index to remain unchanged. One can achieve this by adjusting the divisor, from 10 to 11, such that the index is still 1.1 / 11 = 10% at *t2*. From *t2* onward, we would then keep the divisor as 11 unless the composition of topics changes again, and changes in the discoverability score for the new topic would affect the overall index. For example, in a future time period *t3*, if the observed percentage of harmful results for the added topic increases from 20% to 40%, the overall index would increase from 10%. If at *t4* that topic's score declines to 10%, the overall index would decrease below 10%. This adjusted divisor enables the comparability of metrics over time (from *t1* to *t2* to *t3* to *t4*) and is robust to the addition and/or subtraction of topics.

time	Number of topics	Sum over all topics of discoverability score	Divisor	Index
t1	9	0.9	9	0.9 / 9 = 10%
t2	10	0.9 + 0.2	11	1.1 / 11 = 10%
t3	10	0.9 + 0.4	11	1.3 / 11 = 11.8%
t4	10	0.9 + 0.1	11	1 / 11 = 9.1%