THE JUSTICE COLLABORATORY

YALE LAW SCHOOL

Report Of The Facebook Data Transparency Advisory Group

April 2019

Ben Bradford
Professor, Department of Security and Crime Science,
University College London

Florian Grisel
Research Fellow, Centre National de la Recherche Scientifique and
Reader in Transnational Law, King's College London

Tracey L. Meares
Walton Hale Hamilton Professor of Law, Yale Law School

Emily Owens
Professor, Department of Criminology, Law and Society, and
Department of Economics, University of California, Irvine

Baron L. Pineda
Professor of Anthropology and Latin American Studies,
Oberlin College

Jacob N. Shapiro
Professor of Politics and International Affairs, Princeton University

Tom R. Tyler
Macklin Fleming Professor of Law and Professor of Psychology,
Yale Law School

Danieli Evans Peterman
Research Scholar in Law, The Justice Collaboratory at Yale Law School
Editor and Project Manager

Table of Contents

I. Introduction and Background

Facebook released Version 1 of its Community Standards Enforcement Report ("CSER" or "the Report") in May 2018, and it published Version 2 in November 2018. The CSER shares the metrics that Facebook uses to evaluate its effectiveness in enforcing its "Community Standards." These Community Standards set rules prohibiting certain types of content from being posted on Facebook. Facebook chartered the Data Transparency Advisory Group ("DTAG") in May 2018 to (a) assess Versions 1 and 2 of the CSER, (b) provide recommendations for how to improve its measurement and reporting practices, and (c) produce a public report on its findings.

This report details DTAG's process, findings, and recommendations. It is organized in the following order: Part I.A describes DTAG's charter and process. Part I.B provides a brief background on the CSER, which is discussed in more detail throughout. Part I.C summarizes our main findings and recommendations in response to the three questions within the Group's charter (listed in the following section). Parts II, III, and IV more thoroughly discuss each of the three questions, and explain the basis for our recommendations. Part V concludes with a discussion of steps we believe Facebook could take to improve its reporting on and transparency about Community Standards enforcement.

A. ABOUT THE DTAG

The seven members of DTAG are experts in measurement and the role that metrics play in building legitimate, accountable institutions. The members of DTAG have a range disciplinary backgrounds, including anthropology, economics, law, political science, psychology, and sociology. DTAG is co-chaired by Tracey L. Meares and Tom R. Tyler, who are faculty co-directors of The Justice Collaboratory at Yale Law School.[1] The members of DTAG are:

Ben Bradford
Professor, Department of Security and Crime Science, University College London

Florian Grisel
Research Fellow, Centre National de la Recherche Scientifique and Reader in Transnational Law, King's College London

Tracey L. Meares
Walton Hale Hamilton Professor of Law, Yale Law School

Emily Owens
Professor, Department of Criminology, Law and Society, and Department of Economics, University of California, Irvine

Baron L. Pineda
Professor of Anthropology and Latin American Studies, Oberlin College

Jacob N. Shapiro
Professor of Politics and International Affairs, Princeton University

Tom R. Tyler
Macklin Fleming Professor of Law and Professor of Psychology, Yale Law School

Danieli Evans Peterman
Research Scholar in Law, The Justice Collaboratory at Yale Law School
Editor and Project Manager

[1] The Justice Collaboratory is an interdisciplinary group of scholars who work on translating social science research into evidence-based social policy and institutional design, with a particular focus on public trust in authorities and institutions. Much of the Justice Collaboratory's theoretical work applies to social media platforms, as relations between social media platforms and members of online communities resemble relations between governments and members of 'real-life' communities.

The DTAG's Charter

At the outset, Facebook and DTAG agreed on three specific questions within the scope of DTAG's review:

1. Is Facebook accurately identifying content, behavior, and accounts that violate the Community Standards policy, as written?

2. Assuming violations are being counted accurately (question 1), are the publicly-released metrics the most informative way to categorize and measure those violations as well as Facebook's response to them?

3. The current metrics are focused on Facebook's efficacy in enforcing its standards, as written. What additional information does the public need in order to be able to understand Facebook's content standards enforcement efforts and to evaluate the legitimacy of those policies?

Questions beyond the scope of DTAG's Review

DTAG's review was limited to the three questions above. DTAG was not tasked with evaluating any of the following:

- The content of the Community Standards. For example, we were not asked to evaluate whether or not Facebook's definition of hate speech comports with legal or commonly understood definitions of hate speech. Rather, we reviewed whether Facebook is using statistics that are accurate and meaningful for describing its enforcement Community Standards on hate speech, as written.
- Facebook's policies on collecting, storing, using, and sharing users' data.
- Facebook's policies and practices surrounding advertising, including how Facebook counts the number of views ads receive, how Facebook markets ads to buyers, and how Facebook targets ads to users.
- Facebook's policies with respect to "fake news" or misinformation, as neither of these categories were counted as violations within the first two versions of the Community Standards Enforcement Report.

None of our conclusions should be taken as an assessment or endorsement of any of the policies listed above, or any of Facebook's other policies, except for the three specific questions within the scope of DTAG's Charter.

Facebook agreed to provide DTAG with the information necessary to assess the three questions in its Charter. To facilitate our work, we entered into non-disclosure agreements with Facebook, which allow us to review internal information about Facebook's processes used to generate its enforcement data, estimation approaches including underlying assumptions, and technical constraints related to measurement of its Community Standards enforcement. We received a

series of briefings from Facebook over the course of six months, from July through December 2018. We met with a number of teams at Facebook, including the team that writes the Community Standards; the team that trains, manages, and audits content reviewers; the team that designs automated screening technology; and the team that measures violations of the Community Standards and Facebook's responses to them. The parties agreed that DTAG's public report would not include any confidential information—i.e., information about Facebook's business and its products that Facebook or other sources have not previously made public.

While we talked to a large number of engineers who design and oversee content enforcement and measurement systems, we did not speak directly with engineers maintaining systems day-to-day. Hence we cannot evaluate the extent that Facebook's daily operations deviate from the process that was described to us in oral briefings. We reviewed one document, produced in response to our questions, which detailed Facebook's equations for sampling content on the site and calculating its prevalence metric (discussed in more detail below). However, we did not seek to audit Facebook systems or to review code that instantiates the procedures and algorithms described to us. Our conclusions are based on the assumption that Facebook actually implements policies and procedures in the manner they were described to us, and that the information Facebook shared on performance-specific tasks is accurate. To the extent these assumptions are false, our conclusions do not apply.

Compensation

As is standard for technology industry scientific advisory boards, the members of DTAG also received financial compensation in the form of a pre-determined, fixed honorarium, paid prior to our assessment and reporting. DTAG members are not Facebook employees. Our compensation was paid in full before the report was written, and as such, it is not tied to any conclusions, assessments, or recommendations that we provide Facebook. Decisions about whether and how to implement our recommendations will be made by solely by Facebook.

B. FACEBOOK'S COMMUNITY STANDARDS ENFORCEMENT REPORT

Facebook's Community Standards define the content that is prohibited on Facebook. The Community Standards, in addition to the internal guidelines that are used for interpreting the standards, are available at: https://www.facebook.com/communitystandards. The preamble to the Community Standards states they are rooted in three principles: safety, voice, and equity. The Standards identify five general categories of regulated content, each of which encompass more specific categories of violations: violence and criminal behavior (e.g., credible threats of violence, terrorist propaganda, selling illegal goods); safety threats (e.g., child nudity, sexual exploitation, bullying and harassment); objectionable content (e.g., adult nudity, hate speech); integrity and inauthenticity (e.g., fake accounts, spam); and intellectual property violations. The preamble does not explain the process by which those principles were translated into a concrete set of rules nor how the company plans to develop them over time.

In order to identify content that violates the Standards, Facebook uses a combination of

automated and human review. This enforcement process and Facebook's method of auditing it are detailed in Part II, which assesses the accuracy of Facebook's Community Standards enforcement.

In May 2018 Facebook released Version 1 (V1) of CSER. It released Version 2 (V2) in November 2018. Both versions of the CSER are available at: https://transparency.facebook.com/community-standards-enforcement. At the same time, Facebook published a separate guide, *Understanding the Community Standards Enforcement Report*, which gives more detail about the methodology underlying the report. It is also available at the same link as the CSER. In *Understanding the Community Standards Enforcement Report*, Facebook stated "[i]n recent years, we have begun focusing more on measuring and labeling these efforts cleanly and consistently so that we can regularly publish reports that are meaningful and comparable over time." We take this statement to imply that Facebook intends to publish updated versions of the CSER "regularly" in the future, in order to update the public about how well it is doing at identifying and removing violating content, and to document how this changes over time. As of April 1, 2019, Facebook has not published a third version of the report.

The first two versions of CSER include three metrics:

1. ***Prevalence***: The "estimated percentage of total views that were of violating content. (A view happens any time a piece of content appears on a user's screen.)" This estimate is based on a human review of a stratified random sample of all content views on Facebook at a given point in time, regardless of whether or not that content was ever flagged by users, detected by Facebook, or acted on by Facebook.

2. ***Actioned content***: "The number of pieces of content (such as posts, photos, videos or comments) or accounts we take action on for going against standards. 'Taking action' could include removing a piece of content from Facebook, covering photos or videos that may be disturbing to some audiences with a warning, or disabling accounts."

3. ***Proactively Actioned Content***: "The percentage of all content or accounts acted on that Facebook identified before users reported them. (The remaining percentage reflects content and accounts Facebook took action on because users report them to us first.)" This metric is calculated as "the number of pieces of content acted on that [Facebook] found and flagged before people using Facebook reported them, divided by the total number of pieces of content [Facebook] took action on."

These metrics are reported separately for different categories of violations. Some categories of violations were added in CSER V2, so their metrics are not included in V1. Table 1 lists the types of violations, and identifies the versions of the CSER where each of the metrics are available.

Table 1: Versions of CSER in Which Metrics were Available for Each Violation Type

Violation Type	Prevalence	Actioned Content	Proactivity
Adult nudity and sexual activity	V1 & V2	V1 & V2	V1 & V2
Graphic violence	V1 & V2	V1 & V2	V1 & V2
Fake accounts	V1 & V2	V1 & V2	V1 & V2
Global terrorism	Not available	V1 & V2	V1 & V2
Hate speech	Not available	V1 & V2	V1 & V2
Spam	Not available	V1 & V2	V1 & V2
Bullying and harassment	Not available	V2	V2
Child nudity and sexual exploitation	Not available	V2	V2

Note: "V1" indicates that the metric is available in Version 1 of CSER; "V2" indicates that the metric is available in Version 2 of CSER; "Not available" means not available in either Version 1 or Version 2 of CSER.

Source: Versions 1 and 2 of the CSER.

For some types of violations, there is no prevalence estimate provided in either V1 or V2. This is because Facebook felt that it was unable to estimate these metrics reliably given the state of their internal systems (e.g. performance of models used to conduct stratified sampling for a given violation type). Facebook teams explained how it came to these determinations around prioritization as well as the thinking behind their future development path for the CSER. For example, because the vast majority of global terrorism is caught before it receives and views (this process is detailed below), a prevalence estimate for this category (measured in terms of number of views) would not be particularly informative, and it is a lower priority. For bullying and harassment, it is difficult to detect these violations unless a user reports them, since they depend on how a user subjectively understands the communication in question. It is difficult to independently sample and label these violations that depend on a user's subjective understanding. For spam, it is difficult for humans to sample content and detect spam, because some spam is defined by the source and distribution methods, which cannot be determined by human content review alone. We found Facebook's reasoning on such matters to be generally sound, though would suggest more transparency on the roadmap and on the reasons why some violation types were easier to assess in the first two reports.

In addition to the metrics listed above, V2 of the CSER states that Facebook has two other metrics "under development":

1. ***Speed***: "How quickly did we act on violations?"
2. **Accuracy**: "How often did we correct mistakes?"

As of April 1, 2019, Facebook had not publicized a timeline of when it plans to release these metrics. We recommend Facebook be more transparent about when it plans to release metrics under development.

C. SUMMARY OF FINDINGS AND RECOMMENDATIONS

This section summarizes DTAG's main recommendations with respect to each of the three questions within our charter. In Parts II, III, and IV, we elaborate on our findings and the basis for our recommendations.

> **QUESTION 1:**
>
> **Is FB accurately identifying content, behavior, and accounts that violate the Community Standards policy, as written?**

Overall, Facebook's system for enforcing the Community Standards, and its methods of auditing the accuracy of that system, seem well designed. Our assessment on this point is limited to the process Facebook follows, according to Facebook's description of it. We did not conduct a sufficiently detailed audit to assess how accurately Facebook implements this process, or whether the performance reported in the CSER is accurate. Facebook's approach to identifying violations combines human annotation and machine learning in a sensible manner. Their audit procedures are reasonable given the volume of material in question and, as described to us, should provide an unbiased estimate of enforcement accuracy. Overall, the Facebook system is a dynamic one in which the results of its audits are fed back to modify protocols and provide feedback to its content reviewers. Facebook is making an effort to improve accuracy through this process, as demonstrated by several changes from CSER V1 to V2. However, we find there are several ways in which Facebook could be more transparent to the public about the accuracy of its Community Standards enforcement systems, and could provide users with additional helpful information. We also suggest Facebook do more to involve users in the process of enforcing the Community Standards.

In response to Question 1, our main recommendations are:

- Prioritize releasing accuracy rates for both human and automated decisions.
  - For review of human decisions, which are subject to an intensive review by a second panel of human reviewers, we advise against attempting to calculate separate error rates for clear violations and ambiguous cases.

- For error rates from automated decisions, we recommend releasing the false positive, true positive, and false negative rates, as well as precision and recall.

- Release reversal rates from appeals separately. Rates of reversal on appeal should be made public, but they should not stand in as the sole public metric of accuracy.
- Share statistics on human reviewers' inter-rater reliability, which differs from measures of accuracy calculated via the review of human decisions referenced above.
- Check reviewers' judgments not only against an internal 'correct' interpretation of the Standards, but also against a survey of users' interpretations of the Standards.

QUESTION 2:

Assuming violations are being counted accurately (question 1), are the publicly-released metrics the most informative way to categorize and measure those violations as well as Facebook's response to them?

The metrics released in V1 and V2 of the CSER are reasonable ways of measuring the values that Facebook is trying to measure. They are analogous to metrics of crime currently published by governmental agencies, as described below. Between V1 and V2 of the CSER, Facebook made several important changes in how it counts content, and this made the metrics even more accurate. The improvements which Facebook reported were in process for future reports are sensible and will, if completed, provide additional valuable transparency. Nonetheless, we recommend several ways of improving the metrics currently in the CSER, and we recommend releasing additional metrics that would help contextualize the current ones. We note that Facebook's measurement efforts are complicated by the fact that it is working with immense volumes of data, on systems not designed for this measurement task. As a result there may be substantial engineering effort involved in responding to these recommendations, which we are unable to quantify.

In response to Question 2, our main recommendations are:

- Report prevalence two ways: (1) Number of violating posts as a proportion of the total number of posts; (2) Number of views of violating posts as a proportion of all views. In V1 and V2 of the CSER Facebook reported only the second metric.
- Explore ways of relating prevalence metrics to real-world harm. E.g., Is an increase in prevalence of hate speech posts correlated with an increase in ethnic violence in the region; or an increase in removals of hate speech posts correlated with a decrease in ethnic violence?
- Explore ways of accounting for the seriousness of a violation in the prevalence metrics. Global terrorism content, for example, may or may not include graphic violent imagery.
- Report prevalence measures in sub-populations, e.g., specific geographic regions or languages.

- Break out actioned content measures by type of action taken (e.g., content taken down, content covered with warning, account disabled).
- Report actioned content as a proportion of total estimated violating content.
- Explore ways of accounting for changes in the Standards and changes in technology when reporting metrics in CSER.

Question 3:

The current metrics are focused on FB's efficacy in enforcing its standards, as written. What additional information does the public need in order to be able to understand FB's content standards enforcement efforts in order to evaluate the legitimacy of those policies?

While the CSER is a positive step toward transparency, we make a number of suggestions about additional measures Facebook could take to enhance the public's understanding of how it regulates content on the platform. This, in turn, could help build public trust and legitimacy.

In response to Question 3, our main recommendations are:

- Explore ways to enhance bottom-up (as opposed to top-down) governance. These models are described in more depth in Part IV.A.
- We identify a number of specific ways Facebook could build elements of procedural justice (participation and voice, fairness, conveying trustworthy motives, treating people with respect and dignity) into its process for Community Standards enforcement.
- For the sake of transparency, we recommend Facebook explore ways of releasing anonymized and aggregated versions of the data upon which the metrics in the CSER are based. This would allow external researchers to verify Facebook's representations.
- We identify a number of specific ways Facebook could modify the formatting, presentation, and text of the CSER documents to make them more accessible and intelligible to readers.

The remainder of this report proceeds in the following order: Part II discusses our findings and recommendations in response to the Question 1 of our Charter; Part III covers Question 2; and Part IV covers Question 3. Part V concludes with a summary of the steps we believe Facebook could take to improve its reporting on Community Standards enforcement.

II. Is Facebook accurately identifying content, behavior, and accounts that violate the Community Standards policy, as written?

A. THE COMMUNITY STANDARDS ENFORCEMENT PROCESS

Facebook's process for identifying and evaluating potential violations of the Community Standards involves several phases of algorithmic and human review. The following paragraphs describe the different steps in Facebook's process of detecting potential violations of the Community Standards, and evaluating whether questionable content actually violates them:

- **Proactive match and action**: Content that is submitted to Facebook is immediately screened by an automated process. This automated screening proactively identifies and blocks any content that matches certain content that Facebook has previously removed as terrorism-related imagery or child exploitation. For example, if one specific image of terrorism-related imagery has been removed from Facebook in the past, when a user tries to post that same image, it is proactively matched to this known illicit image, and automatically blocked. Proactive match and action also detects and removes content that matches some previously-identified spam violations. This preliminary automated screening does not check new posts against every single previously-identified violation. Doing this would result in a long delay between when the user posts content and when the content appears on the site. Hence, at the proactive matching phase, Facebook focuses on terrorism-related imagery and child exploitation—these images are both very harmful and amenable to proactive matching. Pieces of content that pass this preliminary screening are displayed on the site. Once content is displayed on the site, there are two further methods of monitoring for violations: automated detection and human reports.

- **Automated detection**: A different set of algorithms screen content that has been posted (meaning it passed the first level review described in the preceding paragraph) to identify potential violations.[2] These algorithms assess content for similarity to a large number of specific patterns—i.e., images, words, and behaviors that are associated with different types of violations. Here it is important to note that detecting violations is not always black and white. Certain words or images might be associated with violations, but this does not mean that all posts containing those words or images are violations. For example, a post containing an image of a nipple is a violation if the image is pornography, but it is not a violation if the image is a piece of painted artwork or a mother breastfeeding. For this reason, the image of a nipple alone is not enough to classify a post as violating. Other factors, such as the rest of the photograph, the identity of the person posting it, the comments, likes, and shares, all shed light on the content's meaning. Algorithms take many of these features into account in order to calculate the likelihood that any given piece of content violates a Community Standard.

[2] The classifiers are different for different types of violations (for example, behavioral patterns that indicate a fake account are different from images that indicate graphic violence, which are different from the images that indicate nudity).

The list of classifiers is regularly updated and the algorithms are retrained to include information that is learned as more violations are identified (or overlooked). Note that this continual evolution of classifiers, as well as the level at which content is deemed "likely" to violate, varies over time. Both of these variables will directly impact the proactivity rate reported in the CSER. These variables are not directly related to how much violating content actually exists on Facebook, but may influence the prevalence rate as changes causing more violating content to be proactively removed will mean that less such content is shared and viewed, thereby reducing the prevalence rate as currently calculated.

If these algorithms determine that a piece of content is clearly in violation, they may remove it automatically, without human review. However, when these algorithms determine that a piece of content is potentially a violation, but the judgment is uncertain, the content is routed to a human reviewer, who will determine whether it violates the Standards.

- **User reports**: The second way that Facebook identifies potentially-violating content on the platform is by users. Users have an option to flag content if they believe it violates the Standards. For instance, on Facebook for mobile phones, users can flag a post by clicking on ellipses that appear in the upper right hand corner of any post. After clicking on these ellipses, users are given a menu of options. One option is "give feedback on this post." By selecting this option, a user can report the post as a violation of the Standards, and identify the type of violation. When a user reports content, it is routed through an automated system that determines how it should be reviewed. If this automated system determines that the content is clearly a violation, then it may be automatically removed. If the system is uncertain about whether the content is a violation, the content is routed to a human reviewer.

- **Human reviewers**: Content that has been flagged (by either humans or automated review) as potentially in violation of the Standards is routed to a content reviewer, who is selected based on language, location, and type of violation. Each content reviewer has a que of reported posts to evaluate one by one. Sometimes reviewers can evaluate a post in isolation to determine whether it violates the Standards—for example, an image of graphic violence. But other times the context determines whether a post violates the Standards. For instance, a word that has historically been used as a racial slur might be shared as hate speech by one person but can be a form of self-empowerment if used by another. Reviewers evaluate the context of the post, including the comments and the identity of the poster, in an attempt to assess its intent and meaning.

  In a public post in Facebook's Newsroom, originally posted in July 2018 but updated in December 2018, Facebook reported that it maintains a team of approximately 15,000 human reviewers located at 20 different sites around the world.[3] These reviewers were a mix of full-time employees and subcontractors.[4] Reviewers speak over 50 languages. They are trained

[3] Ellen Silver, Hard Questions: Who Reviews Objectionable Content on Facebook—And is the Company Doing Enough to Support Them?, Facebook Newsroom (July 26, 2018) (updated on Dec. 4, 2018 to revise number of content reviewers originally reported), https://newsroom.fb.com/news/2018/07/hard-questions-content-reviewers/.

[4] Id.

in the Community Standards and the more specific enforcement guidance. They also receive hands-on practice making decisions alongside an instructor.[5] According to Facebook, after this training, reviewers receive reports auditing their consistency and accuracy, and identifying areas where they need more practice.[6]

According to this public post in Facebook's Newsroom, content reviewers are not given a quota of the number of posts they are expected to review in a given time period.[7] Facebook states: "content reviewers aren't required to evaluate any set number of posts — after all nudity is typically very easy to establish and can be reviewed within seconds, whereas something like impersonation could take much longer to confirm. We provide general guidelines for how long we think it might take to review different types of content to make sure that we have the staffing we need, but we encourage reviewers to take the time they need."[8] Various press reports suggest that the firms Facebook contracts with to conduct human review sometimes impose quotas and/or have compensation systems that create perverse incentives.[9] Facebook did not provide DTAG with specific information about its contracts with content review firms. It also did not share details about the compensation and incentive structure for content reviewers. Hence DTAG cannot draw any conclusions about these policies.

Auditing reviewers' accuracy: Facebook described to us the process it uses to evaluate the accuracy of human reviewers. Facebook samples all decisions made by human reviewers, and these decisions are re-labeled by a separate panel of three content reviewers. The reviewing panel is given more information about the context of the post, including some additional details about the history of the user who posted the content and the user who reported the content. The reviewing panel is tasked with determining the 'correct' answer in relation to two levels of policy: Standards and protocols. Comparisons to standards involve the core principles that guide Facebook's efforts to manage content. Protocols are the specific implementation rules outlined in the manuals used by content reviewers. Second level reviewers' reach the "correct" judgment based on these considerations, and their judgment can then be compared to the answer obtained during the original reviewers' judgments, in theory to determine accuracy.

Facebook declined our request to share the error rates calculated via this review process. It told us that it has not yet determined how to calculate error rates, because it is uncertain about how

[5] Id.

[6] Id.

[7] Id.

[8] Id.

[9] See, e.g., Casey Newton, The Trauma Floor, The Verge (Feb. 25, 2019), https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona; James Vincent, AI Won't Relieve the Misery of Facebook's Human Moderators, The Verge (Feb. 27, 2019), https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms; Adrian Chen, The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed, Wired (Oct. 23, 2014), https://www.wired.com/2014/10/content-moderation/. For Facebook's response to some of these reports, see Justin Osofsky, Our Commitment to Content Reviewers, Facebook Newsroom (Feb. 25, 2019), https://newsroom.fb.com/news/2019/02/commitment-to-content-reviewers/.

to deal with ambiguous cases. In briefings, Facebook explained that it considers "clear errors" different from errors in cases that are ambiguous—meaning there is not an obvious correct answer on whether the content violates the Community Standards. For reasons discussed in Part II.B, we advise Facebook against distinguishing between errors in cases that are clear and in cases that are ambiguous for the purpose of calculating accuracy.

Auditing the accuracy of automation: To audit the accuracy of content actions taken by automation, Facebook calculates the following metrics:

> Precision: the percentage of posts that were correctly labeled as violations out of all posts that were labeled as violations. In other words, of all the posts that were initially labeled as violations, what percent were actually violations?
>
> Recall: the percentage of posts that were correctly labeled as violations, out of all the posts that are actually violations. In other words, of all the posts that actually violate the standards, what percent were labeled as violations?
>
> Precision and recall are calculated by the following equations:

> Precision = TP / (TP+FP)
>
> Recall = TP/ (TP+FN)
>
> Where:
>
> True Positive (TP) = posts that were labeled as violating the Standards, and actually are violations.
>
> False Positive (FP) = posts that were labeled as violating the Standards, but actually are not violations.
>
> False Negative (FN) = posts that were labeled as not violating the Standards, but actually are violations.

Appeals/review system: Separate from the process for auditing the accuracy of human reviewers, Facebook has recently instituted an appeals or review system. For certain types of violations, users whose posts have been deemed a violation are given the option of seeking review of that decision. Review is not available for certain types of content that are seen as being especially harmful, e.g., safety and security violations or child pornography. When a user elects review, a different human content reviewer takes a second look at the post. The second reviewer is at the same level as the first one; they are not a higher-level content reviewer with more training or expertise. In this sense the term appeal may be misleading. Review is perhaps more accurate. If the second reviewer determines that the content is not a violation, it will be restored to the site.

Presently, appeals are available only for users whose posts are taken down. They are not offered when a user flags a post as violating, but Facebook decides to leave the post up. Facebook indicated that it plans to make appeals available in these circumstances. We advise Facebook

to do this. Even if posts containing hateful and derogatory content do not technically violate the Community Standards, they are nonetheless terribly offensive and harmful.[10] Users who report these posts may experience significant emotional trauma as a result of them, which might be exacerbated when Facebook declines to take action. Even if the initial decision to leave the post up was technically correct in terms of the Standards, allowing users to appeal these decisions signals that Facebook recognizes the potential harm associated with false negatives (an erroneous decision that the content is not a violation).

B. EVALUATION OF ACCURACY

While Facebook provided the Group with a description of its process of assessing accuracy, it told us that it has yet to develop a metric of error calculated via its process for reviewing human decisions. It also declined our request to provide the TP, FP, and FN, or precision and recall rates for the algorithms detecting different types of Community Standards violations. Facebook explained that it calculates precision and recall separately for each classifier in the algorithm, and it would not be possible to aggregate these rates into one overall precision and recall rate for each type of violation. While we agree it would not be possible absent strong assumptions, we encourage Facebook to choose a set of assumptions which enable it to develop an overall metric.

The Group also did not review the specific lists of classifiers that Facebook's algorithms use to detect different types of violations. Facebook did not offer to share the list of classifiers or codes it uses because, it maintained, it would have been impractical for the group to meaningfully review them, since there are a large number of classifiers which are constantly changing. Facebook also did not share with us an estimate of reversal rates via the appeals/review process.

Because we have not reviewed the code Facebook uses to implement its automated review, our assessment is limited to the process Facebook is using to evaluate the accuracy of its enforcement system, based solely on Facebook's description of that process. We can draw no conclusions about whether Facebook actually follows this process in practice. Also, because we have not seen error rates or reversal rates, we cannot draw conclusions about how accurate the human reviewers and automated decisions actually are.

Generally, based on the information provided to DTAG, Facebook's auditing system for human reviewers does seem well designed and likely to be helpful in leading to increasing levels of accuracy over time. Assuming the posts are sampled randomly, or through a well-defined stratified random sampling process, then the review process is an appropriate approach to estimating accuracy.

The Facebook system is an ongoing dynamic system in which the results of this procedure are fed back to both modify the protocols and to give feedback to content reviewers. Facebook appears

[10] Ariana Tobin et al., Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Speech to Stay Up, ProPublica (Dec. 28, 2017), https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes. This article describes examples of hateful, derrogatory, and harmful posts that Facebook has deemed permissible under the Community Standards. In one example, the post featured "a photo of a black man missing a tooth and wearing a Kentucky Fried Chicken bucket on his head," with a caption stating "'Yeah, we needs to be spending dat money on food stamps wheres we can gets mo water melen an fried chicken.'"

to be making a serious effort to improve their accuracy through this process. When we probed Facebook about various design choices during the DTAG process, the engineers responsible for key choices had well-thought out and adequately researched responses on why they designed the system as they did (e.g., the method for calculating the 95% confidence interval around their prevalence estimates).

However, we have several recommendations for being more transparent about accuracy.

Our first recommendation is that Facebook to prioritize finding a way to calculate and release error rates for human reviewers, as well as automated detection, for each violation type. These rates will surely fluctuate over time, but it should be possible to publicize a range, or an average, for each violation type. Facebook has indicated it intends to publicize accuracy metrics in the future.

For accuracy rates of human reviewers, evaluated via the panel review process, we advise against attempting to distinguish between clear cases and ambiguous cases. A simple and intelligible metric would count as errors all instances when the original reviewer's decision was contrary to the panel's ultimate judgment on the post, regardless of whether the right outcome was clear or ambiguous. What matters is that the original decision was ultimately deemed incorrect upon review by more experienced moderators operating with more comprehensive information.[11]

For accuracy rates of automated detection, we encourage Facebook to publish not only precision and recall rates, but also the FP, FN, and TP calculations, as they are likely more intelligible to the public.

Our second recommendation is to release reversal rates in addition to the accuracy rates discussed in preceding paragraphs. Facebook has suggested that, in future versions of CSER, it plans to release the rates at which decisions are reversed via its appeal/review system. While we encourage Facebook to release reversal rates, we believe reversal rates should not be offered as the sole metric of accuracy. It is important to also release the error rates listed in the preceding paragraphs, which are derived from independent sampling processes. There are two reasons for this. First, many users will not seek review of the initial decision. Hence, an unknown number of erroneous decisions will not be appealed. Reversal rates cannot, by definition, capture erroneous decisions that were not appealed. Similarly, in the criminal justice system, exoneration rates are not a good measure of mistaken convictions because (a) many people plead guilty to lesser charges (foregoing any right to appeal); (b) some finish their time before the appeals process is complete; and (c) some simply don't appeal. A second reason that Facebook should not use reversal rates as a measure of accuracy is that appeal/review is currently only available when the initial reviewer determines content violates (and not when they determine it is non-violating). This means reversal rates will reflect only one type of error, false positives (erroneous decision that content violates), but not the other important type of error, false negatives (erroneous decision that content does not violate). It is important to report both types of error, since both may be associated with significant harm.

[11] Facebook further distinguishes between ambiguous content and "grey area" content, which does not violate the Community Standards, but is nonetheless problematic. "Grey area" content should not count in a measure of accuracy, since the decision was technically a correct application of the Standards, even if the current Standards are wrong or problematic.

Our third recommendation is that Facebook report the percentage of potentially-violating content that is actioned by machines and the percentage that is actioned by humans.

Our fourth recommendation is to release performance ratings on audits of human reviewers for each language, or each location. If Facebook is not doing so already, we recommend using a capture-recapture method to assess inter-rater reliability and consistency. Inter-rater reliability differs from the accuracy measured by the panel review process for human reviewers, descirbed above, because it measures whether reviewers are reaching consistent judgments, but not whether those judgments are ultimately the 'correct' interpretation of the Standards. The panel review process is evaluating the latter. However, reviewers could have high inter-rater reliability, but consistently reach the wrong interpretation. Releasing measures of inter-rater reliability would show whether the human judgments are consistent. If the training process for content moderators is getting better over time, then this metric should show that judgments are getting more consistent.

Our fifth recommendation is that Facebook make an effort to externally validate its judgments about whether content violates the Standards. It could do this by randomly sampling users within the same region, who speak the same language. These users could be asked to evaluate whether a particular post violates the relevant provisions of the Community Standards and enforcement guidance. The users' judgments could then be compared to content reviewers' judgments. This would be a way of ascertaining how closely content reviewers' interpretations of the Standards agree with users' interpretations of the Standards. We elaborate on this, along with other ideas for enhancing users' participation, in Part III.

III. Assuming violations are being counted accurately (question 1), are the metrics in the Community Standards Enforcement Reports (Versions 1 and 2) the most informative way to categorize and measure those violations as well as Facebook's response to them?

This Part evaluates the three metrics included in the CSER V1 and V2: prevalence, actioned content, and proactivity. Then we discuss how Facebook should deal with changes to the Community Standards over time.

A. PREVALENCE

The prevalence metric is an estimate of the percentage of the total content-views on Facebook that are of violating content.

The prevalence estimate is not based on reported or proactively detected posts. It is an estimate of how many views include violating content calculated based on a sample of all content views on Facebook. Facebook estimates how many views of violating content there are by randomly sampling all content views on Facebook, regardless of whether the content in that post was flagged as a violation. It labels the content within sampled views as violating or not-violating, then counts how many views contain violating content. This process is entirely separate from the process of detecting Community Standards violations. This allows Facebook to estimate how much violating content is viewed on the platform, including content never flagged as a Community Standards violation.

Analogy between Facebook's prevalence measurement and commonly used measures of crime

The question of how prevalent are Community Standards violations on Facebook is analogous to asking how much crime there is in a particular place, with one important caveat. In order to think about how much crime there is in a particular place, it is necessary to identify a temporal denominator – how much crime in a given year or month. There are four ways criminal justice actors and scholars have thought about how much crime is in a particular place in a particular time:

1. First, how many criminal events did law enforcement become aware of? In the US, this is measured using the Uniform Crime Reporting Statistics (available at https://www.fbi.gov/services/cjis/ucr), which capture the number of offenses known to law enforcement for eight types of significant crimes (as reported by approximately 18,000 state and local law enforcement agencies).

2. Second, one might want to know the number of actual criminal incidents, something estimated from the National Crime Victimization Survey (available at https://www.bjs.gov/index.cfm?ty=dcdetail&iid=245), a Census-administered annual survey which asks a representative sample of 135,000 US residents about their experiences with crime. This

number counts how many criminal events took place, regardless of whether the events were reported, or otherwise discovered, by the police and includes a broader set of offenses than the Uniform Crime Reports.

3. Third, how many people were victims of crime? In the United States, this is the estimated "prevalence rate" calculated from the National Crime Victimization Survey. Note that if one person was a victim of more than one crime, the number of victims might be lower than the number of crimes. In addition, if multiple people were the victim of one crime, then the number of victims would be higher than the number of crimes.

4. Fourth, one might ask how many people were involved in committing criminal activity. Some population surveys, including the National Longitudinal Survey of Youth (available at https://www.nlsinfo.org/content/cohorts/nlsy97) or Monitoring the Future (available at http://www.monitoringthefuture.org/), ask about self-reported criminal activity, and scholars also estimate how many people are involved with the criminal justice system by looking at arrest rates in the Uniform Crime Reports, generally making an implicit assumption about the concentration of arrests in a given population.[12]

The most commonly used measure of crime prevalence in the United States is the first method listed above, the number of criminal events known to law enforcement. Currently, Facebook reports prevalence in a way that is most analogous to the third method listed above, the survey-based approach taken by NCVS to estimate how many people are victims of crime (or, in this case, victims of illicit posts on Facebook's platform). Specifically, Facebook's measures prevalence as the fraction of all views of content on Facebook that include violating content. To estimate this quantity it independently assesses a stratified random sample of all views, as described below. Of course, while this measure is a perfectly valid way of thinking about the violating environment, it is different from the type of information the public generally receives about crime, which is either (1) crimes known to law enforcement or (2) a survey-based estimate of total crimes.

It is important to recognize that none of the different ways to measure crime is objectively stronger or more robust, but rather measure slightly different things; and different policy actions might have different effects on the four different measures. This is also true in the case of violating content on Facebook. For example, changes in what content is pushed to users would affect the number of views of violating content, but not necessarily the fraction of violating content posted on Facebook. Facebook interventions aimed at educating users who repeatedly violate standards, however, might directly reduce the fraction of content that violates and/or the number of users who violate the Community Standards, but only indirectly affect the total number of views of violating content.

One justification for the emphasis on views, rather than content, in reporting prevalence is that in an online setting, there are many more direct "victims" of a violation than in typical street crimes. There is intuitive appeal to this, although it invites a complex discussion of how to classify the intensity of victimization across different Facebook user encounters with violating material.

[12] See BRUCE WESTERN, PUNISHMENT AND INEQUALITY IN AMERICA (2006).

Our recommendation is that Facebook report two metrics. One metric is the number of "bad" posts in comparison to the number of total posts. A second is the number of views of "bad" posts in comparison to the total number of views. At this time Facebook reports the second metric. We do not believe that one metric is inherently better or worse than the other. Providing both gives the public more complete information.

Comment on mathematical conceptualization

Facebook estimates the prevalence of violating content using a stratified random sample of all views on the site (because this sample is based on views, it does not capture violating content that is proactively detected and blocked before it is posted). Within the set of views, it oversamples based on how likely it is that a given piece of content violates the Community Standards. A suite of algorithms, developed using machine learning ("ML"), determines the probability that a piece of content violates one of the Community Standards, and the probability that a piece of content is sampled is a discontinuous, increasing, function of this rate. In other words, views are separated into "bins" based on their probability their content is violating, and each bin is sampled at a different rate. A given piece of content is sampled with probability proportional to the number of times it is viewed because each view enters the sample frame as an independent observation. Internal testing shared with the DTAG shows that this stratification results in Facebook having more precise estimates of the rate at which content deemed most likely to violate actually violates while accounting for the number of views content gets.

A team of human reviewers evaluates the content within the sampled views in order to determine both whether or not the content violates the Community Standards, and whether or not the initial human review was accurate (an "error rate"). Facebook calculates an error rate specific to each type of violation. In order to estimate the prevalence of Community Standard violations, Facebook then estimates the total number of views of violating content divided by the total number of views of sampled content (recall that the sampling unit is a view). This average is then scaled by the average error rate for that type of violation.

We noted two practical issues in Facebook's calculation of prevalence. First, as the ML algorithms improve, and in the absence of changes in the nature of content, one would expect an increasingly large fraction of content to be either highly likely or highly unlikely to violate the Community Standards. In order to address practical issues such as reviewer fatigue (to the extent that the ML algorithm identifies more content as highly likely to violate) this means Facebook must dynamically shift its sampling frame to reflect a continuous change in the distribution of ML scores.

Second, the adjustment for reviewer error, as currently calculated, assumes that the probability of reviewer error is independent of the probability that the content is violating. Facebook's adjustment for initial reviewer error could result in a prevalence rate that is too high or too low, depending on whether the error rate is positively or negatively correlated with the probability the content is violating

Why might prevalence vary from report to report

There are five reasons that Facebook's prevalence measures might vary from report to report.

First, there will be classic sampling variation. The size of the reported confidence intervals, which are generated using standard bootstrapping (resampling) methods, will reflect the potential for this sort of variation.

Second, there will be variation in the actual rate of violating posts by Facebook users, which is intuitively the underlying propensity of Facebook users to violate the Standards. Variation in the prevalence of violating content could be driven by a change in the average number of times someone attempts to post violating content to Facebook. In other words, there may be differences in how frequently users post violating content, holding the Standards themselves fixed. Certain triggering events (for example, a contentious sports game, a public controversy, or an act of terrorism) may cause users to post more content that potentially violates the Standards. This will raise prevalence rates "naturally," without any change in Facebook's Community Standards enforcement policies.

Third, there will be variation in how popular or viral violating content is. Variation in the prevalence of violating content can be driven by a change in the number of times that violating content is viewed after being posted. Facebook actions directly affect this metric, as it seeks to remove content as quickly as possible, or "down ranks" content deemed likely to be violating, which makes that content less likely to appear on viewers' screens. Conversely, if Facebook's algorithms mistakenly rate violating content as non-violating, it will appear on more viewers' screens. Users may also affect this is they shift their behavior to posting content in restricted groups which are not viewed by many people, or embed the content in "click bait" or another viral post.

Fourth, there will be variation in what Facebook defines as "violating," or the guidance given to the reviewers. Changes in the technical or practical definition of violating content will lead to differences in the prevalence of content that violates.

Fifth, variation in the prevalence rate may come from variations in the sampling design that are not reflected in the bootstrapped error. Specifically, as the ML algorithms improve, the distribution of content across the sampled "bins" should become more bimodal. This should not affect the calculation of the unadjusted prevalence rate, but may affect the correlation between the error rate and unadjusted prevalence rate across sampling bins. For example, we might expect a higher error rate for content with a more ambiguous rating. As less and less content receives an ambiguous rating, similarly rated content might become more difficult to classify over time, thus increasing the error rate, and thus increasing a negative correlation between predicted probability of violating and error rate. To the extent that the accuracy of Facebook's adjusted prevalence rate assumes no correlation between the error rate and predicted probability of violation, any change in the actual correlation will result in slightly different adjusted prevalence rates.

Taken together these five sources of variance suggest that while broad trends in prevalence as estimated by Facebook are informative, neither Facebook nor its users should make too much of small prevalence fluctuations between reports.

Of these five sources of variation, the first is essentially noise. The second and third reflect differences in user behavior over time, while the fourth and fifth reflect variation in Facebook's decisions. To the extent possible, Facebook should consider offering information that might disentangle the second and third from fourth and fifth. As we will discuss further in section III. D, one way to do this is to highlight any changes in reviewer guidance by estimating two prevalence rates (using the "old" and "new" guidance), similar to the National Crime Victimization Survey estimates of sexual assault in 1992-1993.[13]

Recommendations for additional measures to report

Facebook might explore presenting prevalence estimates that reflect how governments report street crime, specially reporting the fraction of sampled posts that violate their policy, instead of the fraction of sampled views that violate their policy. It might also be helpful to directly and clearly explain this difference between views and posts, referencing the different ways crime is measured—i.e., number of crimes known to law enfocement vs. number of crime victims. Again, while reporting the number of views rather than the amount of content is not inherently "worse" or "wrong," it is different from the type of information the public generally receives about the amount of crime in a community. In general, DTAG recommends providing more information, with a clear explanation for how this information is different to the information commonly provided on street crime, rather than Facebook limiting the amount of information it provides because of concerns that the information might be misinterpreted. In our discussions with Facebook we were impressed with how seriously staff members took this problem. But we also felt Facebook over-weighted the losses that could come from misunderstanding or potential subversive use of reported numbers compared to the legitimacy gains from providing the public with more information.

For reasons discussed further in Section III.D, Facebook should consider back casting to estimate prevalence under different Community Standards definitions. This might require additional human review, which may be costly. Alternately, Facebook could show the distribution of posts that likely violate under different regimes. It could do this by running new algorithms on archived posts, and showing the different prevalence rates using new algorithms. We elaborate on why this would be valuable in Section III.D, where we discuss this recommendation in more depth.

To the extent technologically possible, adjusting prevalence rates to better reflect the sampling design could improve accuracy.

While there is room for improvement in the ways described above, overall, we commend Facebook on its efforts to measure prevalence accurately. Facebook appears to be making continuous efforts to refine and improve its measurement. It is also adding prevalence metrics for more categories of violating content. And we appreciate that Facebook is forthcoming about its inability to come up with reliable prevalence estimates for certain types of violations, either because violations are viewed infrequently (e.g., terrorist propaganda), or because violations are context-specific (e.g., bullying and harassment), or cannot currently be objectively identified across contexts at scale (e.g., hate speech).

[13] See Bureau of Justice Statistics Fact Sheet: National Crime Victim Survey Redesign (Oct. 18, 1995), available at https://www.bjs.gov/content/pub/pdf/REDESFS.PDF.

B. ACTIONED CONTENT

Both versions of CSER reported on "How much content did we take action on?" The metric is defined as follows:

"We measure the number of pieces of content (such as posts, photos, videos or comments) or accounts we take action on for going against standards. "Taking action" could include removing a piece of content from Facebook, covering photos or videos that may be disturbing to some audiences with a warning, or disabling accounts."

Based on conversations the DTAG had with various internal groups in Facebook as well as reviewing public documents, we believe this metric is intended to convey the intensity of Facebook's enforcement efforts. Facebook has made important improvements in how it calculates this metric from V1 to V2 of the CSER, as detailed on page 27 of the *Understanding the Community Standards Enforcement Report* (available for download at: https://transparency.facebook.com/community-standards-enforcement). Specifically, Facebook now counts the number of pieces of content actioned, regardless of whether that action happened through something specific to that piece of content, e.g. an image, or as an action taken in response to a post with multiple pieces of content.

As currently reported we find the metric useful but incomplete. We identified four areas for improvement in reporting on actions taken to enforce the Community Standards.

First, all forms of action are currently considered the same in the report. We suggest that they be broken out by action type, as removing content or covering photos is a very different matter than disabling an account. Indeed, the rate at which specific content is actioned, but accounts are allowed to remain active, can be understood as a measure of the precision of Facebook's Community Standards enforcement efforts and the level of forbearance being exercised to avoid raising the costs for users who mistakenly violate the standards. Both precision and forbearance are important characteristics of a well-functioning regulatory system.

Second, the rate of action should be reported relative to some benchmark. At present, the rates are not normalized in any manner and the number of actions for different types of content are reported without reference to the overall level of that content. We believe the level of action should be reported with two additional normalizations to better enable the public to understand the extent of violations of Community Standards as well as the distribution of violations across users:

1. For reasons discussed in Part II, Facebook should report the share of violating content that is actioned. Right now, the proactivity number discussed below gets at this to some extent, but a more direct approach would be to report the number of actioned pieces of content that violate a standard divided by the estimated prevalence of that type of violation. With CSER Versions 1 and 2, readers cannot calculate actioned content as a proportion of total estimated violations because the current measure of prevalence is reported in terms of views of violating content, not posts of violating content. While variation in the ratio of actioned content to views of violating content is meaningful, it is not as informative as

the estimated percent of violating content that is actioned. Further, such a ratio might be misinterpreted by the public as a percent. For some types of violations, specifically fake accounts, the appropriate denominator might be the fraction of all accounts, or the fraction of accounts brought to Facebook's attention as potentially fake, in a given time period.

2. Facebook should report how much actioned content comes from how many users. Facebook and its community face a fundamentally different challenge in enforcing Community Standards if 90% of the actioned content comes from 0.1% of the users than if it comes from 30%. The concentration of actioned content across users is a potential marker of community health insofar as more concentration means a greater share of users are abiding by the standards. We recognize that calculation of this metric is made more challenging by the proliferation of fake accounts (754 million fake accounts were removed in Q3 of 2018), but in principle content could be linked to accounts in ways that enabled concentration to be measured. Ideally this would be reported by violation-type and by sub-populations (e.g., by region or language), as the extent to which violations are concentrated among a small group of users varies grately.

Third, as the report notes, there are a range of violation-specific reasons why the number of actions can vary from report to report. This is helpful for readers. We believe Facebook could further improve reporting on actions by showing an alternative calculation for violation types where readily-identifiable irregularities can spike the numbers. For instance, the November CSER report offers the following example:

> *If during a cyberattack spammers posted 10 million pieces of spam and we removed all of them, this number would go up by 10 million. However, this spam may not get many views if we remove it quickly enough. In this way the content actioned number can be high, but it wouldn't have much impact on people's experiences on Facebook. After the attack, the content actioned number might significantly decline, but that doesn't mean we've become worse at detecting spam.*

This suggests to us that, for spam, an alternative metric would involve reporting on actions taken on content not proactively removed, or on content not part of an obvious spike due to cyber activity. However, similar to our response to the distinction between "clear" and "ambiguous" errors, we are hesitant to suggest that Facebook make additional subjective decisions (like defining a cyberattack) in calculating performance measures.

Fourth, it is not clear in the CSER why the report only measures actions taken on terrorist propaganda related to ISIS, al-Qaeda and their affiliates, as opposed to other kinds of terrorist propaganda. Commenting on how Facebook defines terrorist propaganda is outside the scope of DTAG's charge, but we recommend that Facebook report on all measures taken against what it defines as terrorist propaganda, rather than a subset of what it takes actions on. Currently the language of the CSER suggests the latter is reported.

C. PROACTIVITY

Facebook's "Proactivity" metric reports how much violating content is found, flagged and actioned by Facebook before users report it. The metric is presented as measure of effectiveness for Facebook's Community Standards enforcement efforts. As such, the underlying intention is to report on how much harm (to those who might otherwise have viewed the content) has been prevented by Facebook's processes.

The metric is simple, and easy to report: the percentage of all violating content or accounts that were that were identified by Facebook, out of all violating content/accounts (and the remainder was identified by users). The CSER reports, for example, that in Q1 2018 95.8% of adult nudity and sexual activity was found and flagged before users reported it. This means that the remaining 4.2% was found and flagged by users before Facebook detected it. The total amount of violating content actioned is reported alongside this percentage.

There are, however, limitations to this measure, and these are clearly noted in *Understanding the Community Standards Enforcement Report*. First, it does not report how many violations were missed altogether. Second, it does not reveal how long it took to detect violations, and how many views there were during this period. Third, it is sensitive to both the reporting behavior of users and the innately subjective nature of decisions made about some forms or examples of content. We agree with Facebook on all three points.

The first point is particularly important. Despite detailing the caveats above, presentation of this metric in the Community Standards report appears premised on the idea that the denominator is all violating content/accounts. Badging this metric as an indicator of effectiveness certainly suggests this – i.e. the motivation seems to be demonstrating how effective Facebook is in spotting violating content. But the denominator is actually all violating content/accounts that Facebook becomes aware of. There is again a direct analogy here with crime data. The volume of crime that police do not become aware of, because it is not reported by the public or identified in some other way, is known as the 'dark figure' of crime. There similarly is a 'dark figure' of violations that is not represented in this metric.

However, Facebook is in a position to identify this dark figure. It could do so with a process similar to the one it uses to construct the views-based prevalence metric in V1 and V2 of the CSER: It could independently sample all content posted on the platform, including content that has not been flagged as a violation, and count the number of violations within this sample. Reporting actioned content as a proportion of estimated violating content would improve transparency, and be a better measure of effectiveness, since it conveys how much content Facebook is addressing relative to the amount on the platform.

It is also worth noting that the actioned content metric, as currently constituted, is as much a measure of regulatory activity as it is of harm reduction or effectiveness. The particular measure reported by Facebook is most similar to a statistic sometimes used internally by police departments, but not commonly distributed to the public; the fraction of officer-citizen interactions that are initiated by the officer ("on-views") rather than a response to a citizen's call for service.

One implication of this is that, all else equal, Facebook's measure will tend increase or decrease simply as a result of changes in the resources allocated to this process. Moreover, given the complexity of borderline cases – those where it is a judgement call about whether content breaches Community Standards or not – there is also a sense in which increased activity on the part of Facebook creates transgressions, for example when content is removed which no users would ever have actually complained about, or even necessarily found troubling. This is likely to be most pronounced for provisions of the Community Standards that are more nuanced and subtle; nudity, for example, is clearly defined by policy and easy to identify,[14] although the general public may not agree with the policy. Bullying or harassment, by definition, depends on how a user responds to content. One might expect that a higher rate of proactive removal of bullying content would be associated with a higher rate of non-violating content being removed.[15]

The Group also notes that changes in this measure of effectiveness reflect a combination of Facebook proactivity and two different types of user behavior. An increase in the fraction of violating content initially identified by Facebook, as opposed to users, could reflect changes in the Facebook review process, changes in the propensity of users to post violating content, and/or changes in the propensity of viewers to report content. For example, changes in the Facebook interface that encourages users to report violations, or increased awareness of Facebook Community Standards, would tend to make this measure smaller in magnitude (at least in as much as these promote reporting behavior among users). Increased user participation in enforcement of Community Standards should be a policy goal, but it would lead to a reduction in the proactivity metric.

All the points raised above illustrate that there are unavoidable tradeoffs in choosing which metrics to focus on. Different metrics prioritize different goals. A particular metric, like proactivity, will encourage employees to optimize one value. However, in so doing, it might divert effort and attention from other important values. Focusing on a measure of proactivity conveys to employees that it is desirable to increase the percentage of violating content that is proactively removed. If the number of proactive actions taken continues to increase, this could be considered 'good' because Facebook is doing more to catch violating content; but it could also be considered 'bad' because even if the specific metric is 'improving,' either (a) more violating content is being posted, (b) users are not bothering to report violations, or (c) user reported violations are not being actioned, calling into question the effectiveness of the overall regulatory system[16]; and it might also be considered 'counter-productive' because content is being flagged and actioned mistakenly, or when it did not need to be. As more robust time trends become available answers to some these questions will be forthcoming, but the underlying issues will remain. This is not necessarily a call for this metric to be calculated differently, but rather that more care may be needed in presentation and interpretation.

[14] Although there are times when nudity is ambiguous. For example, a shot of a nipple is usually prohibited, but not when the image is a woman breastfeeding or a work of art.

[15] Further, to the extent that proactive removal may suggest to the users that their benign interactions are negative, this may have unintended behavioral effects on users themselves.

[16] See Ariana Tobin et al., Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Speech to Stay Up, ProPublica (Dec. 28, 2017), https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes (discussing examples where Facebook erroneously left up pieces of content reported by users, which Facebook later acknowledged were violations of the Standards).

Recommendations for additional measures to report

Working on the assumption that there is no 'right answer' here, this metric should be accompanied by others that address some of the points raised above, for example:

- A metric that provides an estimate of how much of all violating content is taken down by Facebook (i.e., actioned content /estimated prevalence); and
- A metric that includes a measure of the seriousness of the violations that are actioned/ missed by Facebook. This would allow assessment of whether regulatory activity is correctly calibrated or is instead inappropriately 'sweeping up' very minor forms of transgression.

D. ACCOUNTING FOR CHANGES IN THE POLICY

A key challenge for Facebook in reporting on Community Standards enforcement is dealing with changes in policy. Two particular questions merit consideration:

1. How should Facebook deal with content that violates a new policy, but did not violate at the time of posting?
2. How should Facebook deal with changes in prevalence and other metrics that are a result of policy becoming stricter or more lenient, rather than changes in conduct or in enforcement systems (e.g. improvements in algorithms used to detect violating content or enhanced training of human reviewers)?

The first question has some legal analogy in the form of ex post facto law, which refers to laws which criminalize actions that were legal at the time they were conducted. The United States Constitution prohibits the passage of such laws, as do many constitutions in democratic countries. However, the decision to post is not obviously what Facebook is regulating so much as the content that exists on Facebook's platform. As of February 2019, Facebook stated in the introduction to its Community Standards, these describe "what is and is not allowed on Facebook." In this setting, the issue of ex post facto enforcement can be restated as follows: If FB changes policy and revises algorithm, then runs the new algorithm over all past content, should old content that did not violate at the time it was posted be counted as a violation?

We do not take a stand on whether Facebook should retroactively apply new definitions to old content in terms of removing violating content. While the introduction to the Community Standards suggests that all content, regardless of posting, must satisfy the current Community Standards guidelines, such questions are beyond our remit. The answer must entail some balance between damage done to the community by the existence of the content, the probability the older content will be seen, and the harm to the poster from having the content removed.

As noted above, for purposes of publicly reporting on the extent of violations of the Community

Standards as currently written, we believe that it does make sense to apply current rules to older content. Doing so with respect to prevalence measures in particular will allow the user community to understand trends in posting and also for Facebook and the user community to understand the extent to which users are choosing to abide by current Community Standards.

Our suggestion to apply standards retroactively for purposes of measurement leads naturally to the second question: How should Facebook deal with changes in metrics that result from policy changes as opposed to changes in conduct? Here we propose a simple solution. Facebook should publish a log of significant changes to the Community Standards. Those dates should be laid out in a timeline as part of the periodic CSER releases. Facebook indicated to the Group that it does intend to do this in the near future.

Such documentation will serve two purposes. First, it will allow users to understand whether particular inflection points in prevalence metrics reflect changes in user activity or changes in standards. Second, it will serve to document the evolution of the Community Standards which, while quite dynamic and of recent provenance, are currently presented in Facebook public communications in a static form. Explicitly acknowledging the evolution of the definition of violating content, particularly in response to feedback from users, may serve to enhance the legitimacy of the Standards themselves.

IV. What additional information does the public need in order to be able to understand Facebook's content standards enforcement efforts and to evaluate the legitimacy of those policies?

The preceding Parts assessed how accurately Facebook enforces the Community Standards, and how it measures and reports on this enforcement. This Part discusses more broadly what Facebook could do to make its content moderation practices more transparent and to incorporate users in the process. Doing so might help to build public trust. As described in more detail in Part IV.B, research on trust suggests that people evaluate that trustworthiness of organizations by considering several elements. The first is the degree to which they believe that they are accorded some degree of voice or participation in determining what rules will be or how they will be enforced in evaluating their own posts. The second is the extent to which an organization is transparent and explains what its rules are and how they operate both in general and with respect to particular cases. This involves whether there are opportunities for explanation, dialogue and appeal when there are disagreements about the meaning of particular rules. Finally, people consider whether they feel that their needs and concerns are being taken into account by authorities who are trying to make decisions that are fair to everyone involved.

A. MODELS OF GOVERNANCE

In this section, we analyze the Community Standards Enforcement Reports from the perspective of the broader policy goals that have underpinned their creation, and offer an analysis of how Facebook currently engages in content moderation in support of its Community Standards.

These policy goals were stated in *Understanding the Community Standards Enforcement Report*:

> *We want to protect and respect both expression and personal safety on Facebook. Our goal is to create a safe and welcoming community for the more than 2 billion people who use Facebook around the world, across cultures and perspectives.*
>
> *To help us with that goal, we maintain a detailed set of Community Standards that define what is and isn't allowed on Facebook. We don't allow anything that goes against these standards, and we invest in technology, processes and people to help us act quickly so violations of standards affect as few people as possible.*
>
> *We are sharing the Community Standards Enforcement Preliminary Report publicly for the first time to help people understand how we're doing at enforcing our Community Standards. The report measures how we help to minimize the impact of standard violations on people using Facebook by acting against those violations [...]*

Facebook's immediate goal is informational: as indicated in the last paragraph above, Facebook wishes to improve the public's "understanding" of the ways in which it applies Community Standards by measuring enforcement accuracy. However, Facebook's fundamental goal might be broader, as appears from the first two paragraphs: what Facebook really intends when advertising metrics is to promote "a safe and welcoming community" for its users "across cultures and perspectives."

Beyond the debate concerning the measurement of metrics lies the important question of whether Facebook can achieve its underlying goals. The DTAG, which is composed of lawyers, economists, anthropologists and political scientists, wishes to place the debates on metrics within the broader context of the governance model that Facebook seems to be implementing.

In this section, we analyze how Facebook's governance model has been evolving from a "bottom-up" towards a "top-down" system of governance. We then analyze the effects of this evolution and formulate recommendations.

Bottom-up governance

The traditional governance model on virtual networks is "bottom-up": users define their own norms and actively intervene in their enforcement. Members of the "community" can determine, through relational governance, the norms of conduct that apply on the network. Community governance is found in groups which lack viable legal or governmental authorities, or in which people to not trust those forms of authority. In such settings people manage themselves through shared social norms defined by the group. These norms are "self-enforced." Group members sanction norm breaches by shaming violators and excluding repeat offenders.

This governance model can be illustrated by the system put in place on Wikipedia: The Wikimedia Foundation defines a short set of broad norms that arbitration tribunals (composed of users) are called to apply in case of breach. This model also appears to have had some influence on the normative system put in place by Facebook: Facebook users actively intervene in the enforcement of Community Standards by flagging posts that could be violating these standards (as explained in Section 1); they are also called to provide some "feedback on what they think shouldn't be on Facebook" (*Understanding the Community Standards Enforcement Report*, p. 12). Private "groups" on Facebook can also create their own norms (provided these norms comply with the Community Standards). This influence of "bottom-up" governance is not entirely surprising considering the fact that Facebook was originally conceived as a network of close-knit communities of students.

A bottom-up model of governance presents several advantages. It is cheap, builds a community of users, and usually generates a strong sense of legitimacy. It is noteworthy in this regard that Wikipedia has a broadly positive image (which might also be explained by its "not for profit" nature). On the downside, this governance model is decentralized and does not allow a single entity to build a clear policy agenda. It is also less likely to work effectively in a large community where disagreements might arise concerning the content of applicable norms. For these (and other) reasons, Facebook appears to have moved away from this mode of governance.

Top-down governance

A "top-down" model of governance (also called an "activist model") is one where officials implement a relatively detailed set of rules over a given community.[17] In this governance model, users or citizens have little power over the definition and enforcement of rules, which are placed in the hands of these "officials."

This governance model presents clear advantages. It is usually more effective than a "bottom-up" model at governing a large community. Rules are clear and well-defined, and their application is consistent throughout the community. In addition, the "officials" in charge of defining and applying the rules have expertise, resources and ability to devise long-term policies. In situations in which communities are geographically dispersed or in pluralistic settings where there are few shared norms and perhaps not even a shared sense of community, top down government may be the only viable governance model available.

Facebook's Community Standards enforcement process has features of a "top-down" model. Community Standards have become increasingly long and detailed. Although DTAG was unable to review previous versions of the Community Standards (another reason to publish permanent copies of important documents, as advised in Section IV.C), Facebook stated in *Understanding the Community Standards Enforcement Report* that it "updated" its standards "*with more detailed explanations to help people understand where we draw the line on issues*" (emphasis added). With statements like this, unnamed officials at Facebook are imposing rules on the community. It would be good for Facebook to identify the "we" that is responsible for drawing the line on issues. In addition, Facebook seems engaged in an effort to expand its team of Content Reviewers and to further develop its system of artificial intelligence in order to proactively detect more content violations. We identify below some possible effects of this evolution, and formulate recommendations.

Effects and recommendations

Facebook seems to be currently facing the negative effects of the "top-down" model that it has been implementing. Users feel disempowered in their community. They become increasingly "litigious" concerning the ways in which the standards apply to them. Users also express worries concerning long-term policies that apply without taking their interests into account. They routinely challenge the competence and powers of the "officials" who apply the Community Standards (i.e. the content reviewers and artificial intelligence system). In sum, the network suffers from a crisis of public trust. While it professes its goal to "create a safe and welcoming community," Facebook actively implements a model that is perceived (accurately or not) by users as not taking account of their perspective and consequently weakening the Facebook community. Efforts to create a safe community are conflicting with the desire to create a welcoming community.

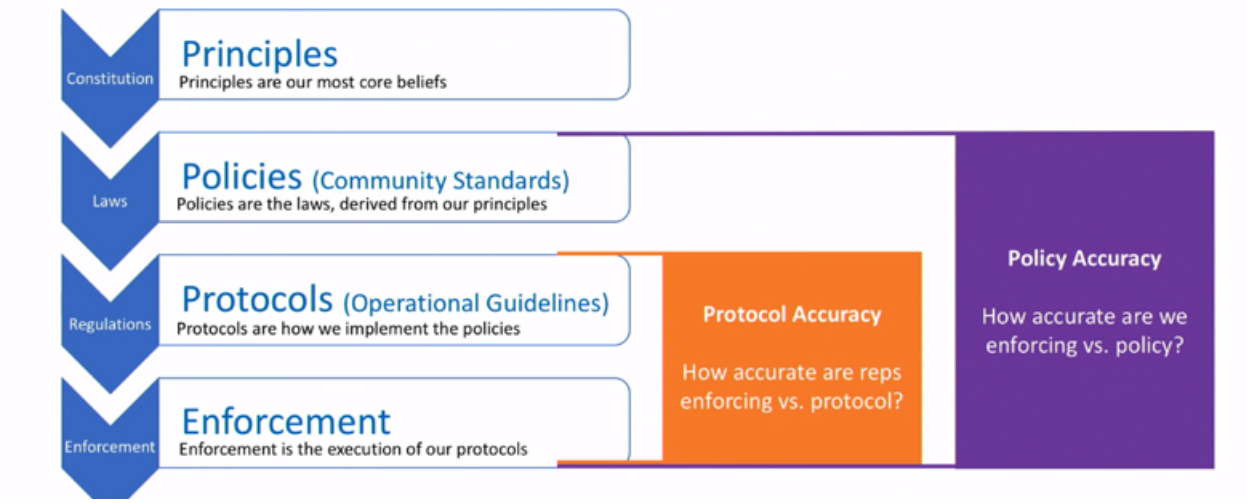
[17] See MIRJAN DAMASKA, FACES OF JUSTICE AND STATE AUTHORITY (1986)

Facebook seems well-aware of this risk, and has responded in a typical "top-down" manner. One of its advertised innovations concerns the creation of an appeal system (and possibly an external oversight board, which Mark Zuckerberg described as being "almost like a Supreme Court").[18] This appeal system will be discussed further in Section IV.B. While the creation of an appeals system can be an effective way to address trust issues (notably by offering "losing" parties an opportunity to elaborate on their reasons justifying their position about their post), this approach is also costly and increases the sense of a gap between users and "officials." Efforts at increasing transparency and providing more detailed reasons in support of decisions (discussed further in Section IV.B), are additional ways to address trust concerns.

We would like to explore the possibility of achieving a better balance between the above-described governance models. We note in this regard that no system is purely "top-down" or "bottom-up": each system presents a blend of "top-down" and "bottom-up" features, and Facebook does in fact present such a blend of features.

However, it seems advisable in the case of Facebook to strengthen its "bottom-up" features in light of its objective to build a sense of community among users. Several presentations to our group began by noting that Facebook's mission is to "give people the power to build community and bring the world closer together" (emphasis added). At the moment, there seems to be a growing contradiction between the stated principles of Facebook (described as its "Constitution" in Figure 1, below) and its protocols and enforcement measures (which entail "Accuracy Measurement" in Figure 1). This is not a comment on the quality of accuracy measurement but upon the impact of the process of content management on the trust that Facebook users have in Facebook.

Figure 1: Slide Showing the Ways Facebook Evaluates Accuracy from Facebook's Presentation to DTAG on December 3, 2018



[18] Ezra Klien, Mark Zuckerberg on Facebook's Hardest Year, and What Comes Next, Vox (Apr. 2, 2018), https://www.vox.com/2018/4/2/17185052/mark-zuckerberg-facebook-interview-fake-news-bots-cambridge.

In line with its stated mission to empower people to build a global community, Facebook could consider a number of recommendations that are outlined below. DTAG is aware that these possibilities may be more or less technically feasible, and that some of them are currently discussed in the public sphere (and probably at Facebook).

- Create an elected "parliament" of users with powers over the definition of Community Standards
- Create a broad partnership with a non-governmental organization specialized in the defense of democratic values and individual liberties.[19]
- Promote the creation of "nested communities" (building on the example of "groups") within which mechanisms of "bottom-up" governance tend to be more effective.[20]
- Organize regular third-party evaluations such as DTAG.

While Facebook has made laudable efforts to advertise the ways in which it enforces Community Standards, these efforts can be described as part of a "top-down" project to build a global, private government that is seeking to regulate content among more than 2 billion individuals. The above recommendations could usefully complement these efforts and improve public perceptions of the company.

B. PROCEDURAL JUSTICE IN INTERACTIONS WITH USERS

We understand that, by releasing the CSER and chartering this Group, Facebook is aiming to build public trust and establish the legitimacy of its efforts to manage content. In this Section, we will discuss ways Facebook could build public trust and legitimacy apart from releasing metrics on its Community Standards enforcement efforts. We identify ways that Facebook could enhance users' trust in the platform by following principles of procedural justice in its interactions with users. We recognize that some of these recommendations may be difficult or even impossible to implement at scale. However, we provide them as goals that Facebook might aspire to, should the technology of the platform allow it.

Background on procedural justice theory

Over the past four decades, a large volume of social psychological research has shown that people are more likely to respect authorities and rules, and to follow those rules and cooperate with those authorities, if they perceive them as being legitimate, i.e. as appropriate people to decide what the rules should be. Perhaps counterintuitively, this research also shows that peoples' judgments about legitimacy do not depend primarily on whether authorities give them favorable outcomes, for example allowing them to post what they want or accepting their arguments

[19] There are some examples of successful partnerships between NGOs and multinational companies, see e.g. the partnership between Chevron and WWF in the 1990s with positive environmental impacts over the Kikori forest in Papu New Guinea.

[20] Nobel Prize Recipient Elinor Ostrom (Economics) advocated the use of « nested enterprises » in order to better regulate large communities.

when they make appeals. Rather, judgments about legitimacy are more strongly swayed by the processes and procedures by which authorities use their authority—namely, whether they adhere to the elements of "procedural justice."[21] Procedural justice is a construct comprising the following four components, which have been shown to have the most influence on perceptions of legitimacy:

- Participation and voice: People consistently report higher levels of satisfaction in encounters with authorities if they have an opportunity to explain their situation and their perspective, and if authorities demonstrate they are listening. This is true even if their participation does not ultimately change the outcome of the authority's decision.
- Fairness and neutrality: People look for indicia that the decisionmaker is being impartial, which is conveyed when the decision is based on generally-applicable rules, and an accurate assessment of the situation. Notable disparities or inconsistencies in enforcement, not explained by any neutral, generally applicable criteria, undermine perceived neutrality. Failure to explain the neutral, generally applicable rules, and explain how they apply to the relevant facts, may also undermine perceived neutrality.
- Respect and dignity: People respond strongly to whether authorities treat them with respect and dignity. This entails treating people like their interests are important and worthy of consideration. Indicia of respect range from politeness to minimizing burdens and impositions on people's time and freedom.
- Trustworthy motives: People are sensitive to whether authorities have benevolent motives. This means that authorities are sincerely trying to act in their best interest, or at least in the best interest of the community. Trustworthy motives are demonstrated by explaining why certain rules and decisions serve the best interests of the individual or the community at large. People are particularly sensitive to any evidence that the decision maker is or is not considering their particular arguments about their situation, their needs or their concerns. Messages which sound impersonal and "boiler plate" in nature do not build trust.

Applying procedural justice theory to Facebook

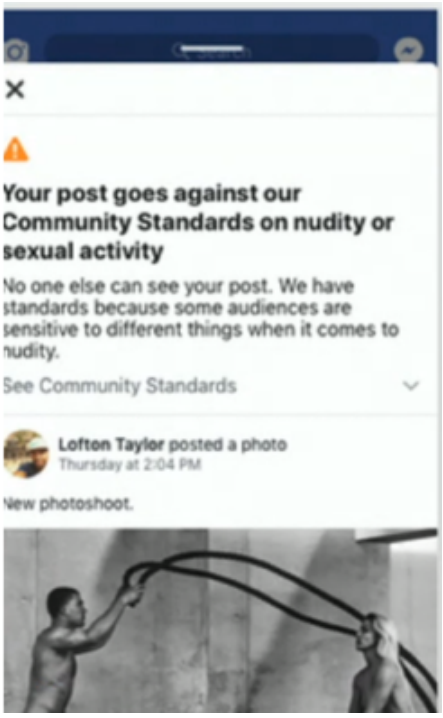
Most research on procedural justice has focused on relationships between members of the public and legal authorities such as the courts and the police. Because Facebook is effectively "policing" content on its platform, the relationship between Facebook and its users is analogous to the relationship between the police and members of the public. Just as we see with individuals and the police, we expect that procedural justice in Facebook's interactions with its users will influence users' judgments about the legitimacy of Facebook's authority, and their acceptance of Facebook's judgments about appropriate and inappropriate content. In the following, we elaborate on three different ways Facebook might enhance procedural justice in dealings with users.

[21] See generally TOM R. TYLER, WHY PEOPLE OBEY THE LAW (2006).

1 Explaining why content does or does not violate the Standards

Figure 2 is an example, provided by Facebook, of the message a user receives when their post has been deemed a violation of the Community Standards.

Figure 2: Message a User Sees When Their Post is Deemed a Violation from Facebook's Briefing to DTAG on December 14, 2018



In this example, the reasons given for blocking the post are vague and cursory. The message appears automated and formulaic, as opposed to an individualized decision. It does not convey the level of detail and nuance of the Community Standards and enforcement guidance (which Facebook has made publicly available). While the message does tell the user that the post has been deemed in violation of the standards on nudity or sexual activity, it does not tell the user exactly what provision of those standards the post violates.

To help the user better understand the basis for the decision, it would be helpful to include the following: (1) identify the specific provision of the nudity and adult sexual content rules that the post allegedly violates, (2) include a link not only to the Community Standards generally, but to the specific provision that was violated here, and (3) include a link to the relevant provision of the enforcement guidance interpreting the Community Standards. Including these details would help users understand why their post violated generally applicable rules, and show this was not an arbitrary or discretionary decision. This might make the decision seem more fair and neutral. (4) To allow voice, the message should also include a clear statement to the effect of "if you

disagree with these rules, we invite you to submit comments in support of revising the Community Standards," and include a link to the page where users can submit these comments. Recall, research shows that having voice is an important predictor of legitimacy, even if one's participation does not change the outcome.

Facebook could also give users a better sense of its motives, and better explain its decisions, by including more details about the purpose behind the rules. The message above includes a very vague statement of the purpose behind the standards—namely that "some audiences are sensitive to different things when it comes to nudity." This statement does not identify examples of audiences for whom nude content may be inappropriate, nor does it explain why this content might be inappropriate for them. The explanation for the rule could be elaborated along the following lines: "The Community Standards prohibit nudity and sexual activity because the Facebook community includes teenagers for whom this content is not age-appropriate. It also includes a people from many different cultures and religions, who have very different values about public displays of nudity and sexual activity. We strive to make sure content is appropriate for everyone in this universal community." Giving these additional details conveys the well-intentioned rationale for the policy, and it might demonstrate trustworthy motives.

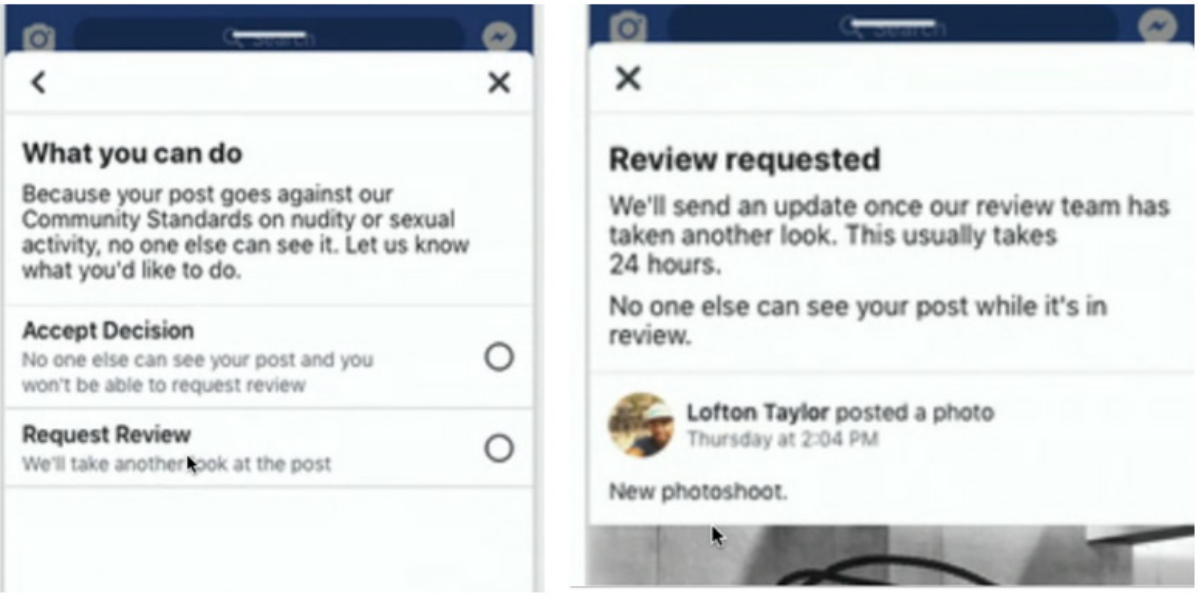
If Facebook does not do so already, it would be advisable to also give an explanation in scenarios where a user reports a post, but the post is left up because it is deemed permissible under the rules. As we mentioned previously, posts containing hateful and derogatory content may not technically violate the Community Standards, but they are still offensive and harmful.[22] Users who have reported these posts may experience significant emotional trauma as a result of them, which might be exacerbated when Facebook declines to take action. Even if it is technically correct in terms of Facebook's standards, the decision to allow this hurtful content may well alienate the user, as it tends to send the message that Facebook sanctions and even endorses the content. To minimize this perception, it would be helpful to give the person who flagged the post an explanation of why the post does not violate the rules. This explanation should include all of the components listed above concerning Facebook's approach to explaining decisions to take down posts: It should identify the specific provision at of the Community Standards and enforcement guidance at issue, and link to these specific provisions. It should also invite users to submit comments in support of revising the Community Standards, and give them a link to this page.

2 Appeal/review decisions

The foregoing applies not only to initial decisions about whether a post violates the Standards, but also to decisions on appeal/review. We discuss appeals/review separately because at this phase there would ideally be even more opportunity to participate, and a more extensive explanation for the decision. Figure 3 shows examples of the interface that offers the user to option of appealing the decision to block their post, which Facebook provided to DTAG.

[22] See Tobin, supra note 10 (discussing an example of content with a racist innuendo that does not technically violate the Community Standards); see also Madeline Varner, et al., What Does Facebook Consider Hate Speech?, ProPublica (Dec. 28, 2017), https://projects.propublica.org/graphics/facebook-hate (displaying pieces of hateful content, some which violates the Standards, and some of which does not).

Figure 3: Interface Offering User the Option of Seeking Review of Facebook's Decision from Facebook's Briefing to DTAG on December 14, 2018



This series of messages does not invite the user to participate or voice their point of view in any way. Voice and participation are important components of procedural justice. Facebook could enhance procedural justice by giving users who choose to appeal the opportunity to write in a brief statement of why they believe the post is acceptable. (To enable users to meaningfully participate, it is necessary to identify the specific provision of the Community Standards that they are charged with violating, as we suggested above.) Such text could be used by the company to better understand how the Standards are understood by users in specific contexts, which would be valuable given the range of cultures represented on the platform.

As of April 1, 2019, Facebook had not provided an example of the text of a written decision on a user's appeal. However, it is important, especially when Facebook rules against the user/appellant, that the decision includes all of the items listed in the preceding section, "Explaining why content does or does not violate the Standards." Ideally appeals decisions would be even more detailed: They would not only refer users to the specific relevant provisions of the Community Standards and enforcement guidance, but they would also include a discussion of how those rules apply to the particular content in question. Furthermore, especially in ambiguous cases where the decision is difficult, it may be advisable to acknowledge the valid arguments that support the user/appellant's case, and explain why those arguments did not ultimately prevail. This would demonstrate that Facebook is making an effort to consider the user's point of view, even when it does not prevail.

Facebook could also give users more indirect participation and voice in the appeals process by recruiting panels of 'juries' comprising randomly selected groups of users. These user juries could be presented with the post that is the subject of the appeal, and the relevant provision(s) of the

Community Standards and enforcement guidance. They would be asked to decide whether the post violates the provision in question. These decisions could either serve as the binding decision upon review (effectively a majoritarian check on reviewers' initial judgments). Or Facebook may be reluctant to give user juries dispositive power, perhaps because doing so would not strike the right balance between top-down and bottom-up regulation (see Part III.A). If so, user juries could be used as an accuracy metric, to measure the extent that reviewers' interpretations of the Standards agree with users' interpretations (see Part II.B) as well as with Facebook's enforcement procedures. Either way, giving users the opportunity to vote on how the Community Standards apply to particular pieces of questionable content would familiarize users with the Standards, and it would give them a sense of having some voice in how they are interpreted. Assuming Facebook actually takes users' views into account (either directly or indirectly), this might help to make the content moderation process seem more democratic.

3 Transparency about the content-moderation process

In addition to the two areas discussed above, there are several other ways Facebook could be more transparent about how it develops and enforces the Community Standards. Possibilities include: (1) Asking users to read and demonstrate knowledge of the Community Standards within a certain period (e.g. one week) of when they first join the site. This could be done by requiring them to answer a few questions about the rules. (2) Publicize a history of when the Community Standards were changed and what prompted the change (see Part III.D). (3) Announce proposed future changes in advance, explain what has prompted the possible revision, and invite users to submit comments on the proposed revision (somewhat analogous to the "notice and comment rulemaking" process used by administrative agencies). (4) Let users know more about the process of selecting, training, and overseeing content reviewers.[23] (5) Publicize error rates (from audits of human reviewers and algorithms), as well as appeal and reversal rates (see Part II.B).

Additional transparency about all steps of the Community Standards development process, and making users feel a part of the process, would help users to understand the basis for the rules and decisions enforcing them. This, in turn will allow users to trust the process more.

Finally, Facebook could enhance transparency by releasing the data used to calculate prevalence. In addition to publishing summary statistics in quarterly reports, Facebook should consider developing a mechanism to make relatively disaggregated data on the prevalence of violating content available to third party researchers. Ideally, the data released would allow the public to replicate some of the prevalence measures reported in the CSER, as well as other references to the volume and regulation of Community Standards violations by Facebook employees or representatives. This is similar to the way the FBI or Census Bureau make aggregations of their data available to the public. It is a non-trivial task to prepare data files for the public in a way that

[23] Posts like the one by Ellen Silver, Hard Questions: Who Reviews Objectionable Content on Facebook—And is the Company Doing Enough to Support Them?, Facebook Newsroom (July 26, 2018), https://newsroom.fb.com/news/2018/07/hard-questions-content-reviewers/, are a good start. However, this may have been seen only by particularly activist users, who either searched for information on this topic, or are in the habit of reading posts in the Newsroom. We suggest Facebook explore ways to more effectively disseminate and publicize these types of communications with users.

protects individual confidentiality, but is a standard task conducted by, for example, data scientists at the Census Bureau.

Providing the underlying data used to calculate the statistics reported in the CSER would be an additional gesture of transparency in itself. It would also allow third party researchers, without any relationship to Facebook, to create reports or analyses of how violating behavior changes in response to policy changes or environmental factors. Because these reports would be done by third parties, the conclusions would almost certainly be viewed by the general public as more credible than internal Facebook research.

As a private company which has a high degree of control over its platform, Facebook has considerable capacity to manage its content regulation process in a top down manner which pays minimal attention to users' views. However, the very existence of the report which we are reviewing highlights the recognition that public views about Facebook and its attitude about the public matter. They matter for the individual user both because disgruntled users find ways to circumvent rules, for example opening multiple accounts. In addition, unhappy customers are less likely to use the site and more likely to seek alternatives to it. There are many reasons that it is important to go beyond simply considering how accurate metrics are and to consider whether the procedures Facebook uses for content management build trust in the company and its leaders.

C. MAKING PUBLIC DOCUMENTS MORE ACCESSIBLE AND COMPREHENSIBLE

In this section we will bracket off the issues of the reliability, accuracy and precision of the metrics used in V1 and V2 of the CSER (which we also refer to as "the Reports"). Instead we will observe and evaluate the ways the Reports are currently presented to the public. In efforts to be transparent, releasing information to the public is a first step, but it is also important to make that information easy to find and easy to comprehend. We will analyze how the CSER and related documents are presented to the public, and we will suggest ways to make these documents more accessible and comprehensible. We identified three ways that Facebook could improve on this front:

1 Make the CSER and related documents easier to access

The CSER are not prominently featured on Facebook's website. In January 2019, when DTAG was evaluating the accessibility of the CSER, if a Facebook user searched in the main search window in Facebook for "Community Standards Enforcement Report," they were presented with outside news coverage of the release of the Community Standards and other Facebook content about Community Standards but the Report itself was not foregrounded. (Of course, this would vary by user.) During the same time period, if one did a Google search for "Facebook Community Standards Enforcement Report," the first two results were Facebook "newsroom" articles introducing the reader to the Report (which can be accessed by clicking on a hyperlink). The first was an essay by Guy Rosen, titled "How Are We Doing at Enforcing Our Community Standards?" dated Nov. 15, 2018. The second was an essay by Guy Rosen from May 2018, "Facebook Publishes Enforcement Numbers for the First Time." Both of these articles provide a clear and engagingly

written summary of the Report and Facebook's motivations for publishing it. They provide a human and personal introduction to the Report itself, which, in accordance with genre conventions for a factual report, has a more sterile tone. Including humanizing introductions like these might be useful for helping readers understand the motivations behind the report, and we do not intend to discourage Facebook from doing so. However, we do advise Facebook to take steps to make the CSER itself more prominent in searches within Facebook, as well as from outside search engines.

On a related note, we would recommend that the present and future Reports be made available in .pdf format in addition to their web-based version. As of April 1, 2019 (as far as we understand) a request for the document would be fulfilled by providing a url for a Facebook web page. This is understandable but it does construct the CSER as a living document that is subject to change at any moment. Given that one of the goals of the CSER is to establish a set of facts it would be advisable to have permanent copies of each Report that are not subject to alteration. This goal is best met with a .pdf format. At present only CSER V1 is available as a .pdf, and this .pdf is accessible only after the user downloads and unzips a .zip file. We recommend making all versions of the CSER available in .pdf format that can be downloaded directly from the website without having to download and unzip a .zip file.

The CSER also includes links to a number of supporting documents. References to supporting documents are helpful, as they give the reader context for the report. However, when there are errors, mistakes, or omissions in the satellite documents, the CSER becomes less credible. For example, in a December 28th "Newsroom" article titled "Facts About Content Review on Facebook" the anonymous Facebook author, in response to an article by the New York Times alleging that Facebook meetings on content review are ad hoc and superficial, writes the following: "Last month we started publishing minutes from these meetings, and early next year we plan to include a change log so that people can track updates to our Community Standards over time." When DTAG reviewed this article in January 2019, there was a hyperlink over "publishing minutes" that lead to an essay by Mark Zuckerberg titled, "A Blueprint for Content Governance and Enforcement." This essay did not include the minutes for the meetings, nor did it include a link to them. It was not clear how the reader could find the minutes that the articles reference. Documents that Facebook references within the CSER and other posts related to Community Standards enforcement should be easy to find. For instance, the link referring to the minutes of Community Standards meetings should not lead to another blog post talking about the minutes of the meetings. It should lead to the minutes themselves. Facebook should apply the same standard of accuracy and rigor to the satellite documents it applies to the CSER.

We acknowledge that Facebook is making efforts to communicate with users via the media (texts, videos, link to televised debates, letters from executives, etc.) that it presents in the "Newsroom" and other Facebook pages. However we think that information about Facebook's policies and practices could be shared with users in more effective ways. The blog posts and letters in the Newsroom are not organized in any systematic way, i.e., according to topic or issue area. They are inconsistent in terms of formality, length, and level of detail. Some are cursory while others are more comprehensive. This makes it difficult to readily access all of the information that Facebook has published regarding a particular topic, such as how content reviewers are supervised and

trained. We recommend creating one webpage with an intuitive heading, such as "Community Standards Enforcement Reports and related documents," that includes links to all of these documents (including the most recent and past versions), organized under specific topic headings (e.g., setting Community Standards, managing and supervising content reviewers, detecting violations of the Community Standards). Under topic headings, users could find briefings that comprehensively describe Facebook's policies and practices related to that particular topic or issue (e.g., a briefing covering how Facebook manages the independent contractors who provide content reviewer services, including specific examples and illustrations). Under each topic area, in addition to comprehensive topic-specific briefs, Facebook could also list links to more informal blog posts and other relevant materials.

2 Provide more transparency about how Facebook approaches the moral, ethical, and legal challenges associated with Community Standards enforcement

We believe Facebook could help users better understand their content moderation policies if it were to (1) demonstrate its awareness of the intellectual, logistic, and moral depth of the Community Standards enforcement effort to the public, (2) put a human face on its efforts, and (3) respond directly and indirectly to the emergent criticisms of the platform. Facebook could help users understand the challenges that go into devising and enforcing the Community Standards by incorporating more vivid examples of the types of borderline violating content that comes across the platform. In particular, sharing examples of "hard cases" akin to ones shown to DTAG, might help users understand the nuance of Community Standards enforcement and some of the judgments that Facebook must make. The statistics in the Report cannot capture the complexity of creating and enforcing Community Standards, nor do they convey the extent that Facebook has grappled with how to best resolve these challenges. For example, a tech blog that briefly covered the release of the Report wrote, "[t]he report is a clear indication of the fact that Facebook is committed to making the platform an amiable place for users. Although the fact that [the] report spurts just numbers and no examples were displayed to back the numbers . . . jeopardizes the credibility of the report."[24] In addition to publicizing examples of "hard cases," Facebook might create a document that explains the reasoning behind the Community Standards—i.e., why it has chosen to draw the line where it has; why it has chosen to revise the Standards when it does so.

3 Present data in a way that is more interactive and user-friendly

We note that CSER V2 added a new drop down list with three categories – standards enforcement, legal requests, and internet disruptions. The reader can then download Facebook data (in the form of excel spreadsheets) related to intellectual property inquiries that Facebook receives, requests from governments for user information that it receives and a summary of the frequency of governmental interference in the availability of Facebook to users across the globe. Producing data like this and including it prominently in the Report is an effective way to show transparency. We realize that this is a new practice but there is room for improvement in how these data are presented. We believe it would be easy for Facebook to convert these

[24] 7 Important Facts From Facebook's Latest Community Standards Report, Fossbytes (May 16, 2018), https://fossbytes.com/important-statistics-facebook-community-standards-enforcement-report/.

excel spreadsheets into formats that convey their messages more evocatively. For example, the spreadsheet that tallies the service interruptions in Facebook across the globe could be presented as an interactive world map, which allows users to look at data for specific geographic regions, and includes both numbers and clickable content that illustrates under what circumstances these events occurred (e.g. declaration of martial law, natural disaster, holidays, etc.). Similarly, the metrics in the CSER might be presented via an interactive map or dashboard that allows users to drill down by country, region, and sub-categories of content within each violation-type.

In considering ways of presenting data about Facebook practices and the environment in which Facebook operates, we can imagine other formats for presenting information. To name one, a relatively new innovation in conservationist and environmental engineering is the environmental "dashboard." These dashboards are publicly located computer screens that provide live demonstrations of real time changes in resource consumption either locally or globally. On a local level, for example, a college might illustrate the real-time water consumption of each dormitory with the goal of encouraging students to take shorter showers and reduce water consumption. Facebook could potentially present the metrics in the CSER in an interactive map or dashboard that allows users to drill down by country, region, and by violation-type. Of course, there are possible complications that could be imagined with implementing a dashboard system but we encourage Facebook to examine practices such as these that are being developing in other contexts. Facebook could also explore ways of presenting information about positive connections made on the platform, in addition to its metrics about negative and harmful communications. For instance, the amount of money raised for charitable causes.

V. Conclusion

We conclude with a brief recap of the steps we believe Facebook could take in order to improve the metrics in CSER and transparency about Community Standards enforcement practices.

1. Release accuracy rates. Accuracy rates for human reviewers should include all errors identified via the panel review process, regardless of whether the error was clear or ambiguous. Accuracy rates for automated reviewers should include FP, FN, and TP rates, as well as precision and recall.

2. Release review/appeal and reversal rates separately. Rates of reversal on appeal should be released, but they should not stand in as the sole public metric of accuracy.

3. Provide information about the percentage of posts that are actioned by automation, and the percentage actioned by humans.

4. Check reviewers' judgments not only against an internal 'correct' interpretation of the Standards, but also against users' interpretations of the Standards.

5. Report prevalence measures not only as a percentage of the total estimated number of views, but also as a percentage of the total estimated number of posts.

6. Explore ways of relating prevalence metrics to real-world harm. E.g., an increase in prevalence of hate speech posts correlated with an increase in ethnic violence in the region; or an increase in removals of hate speech posts correlated with a decrease in ethnic violence.

7. Explore ways of accounting for the seriousness of a violation in the prevalence and proactivity metrics.

8. Report prevalence measures in sub-populations, e.g., specific geographic regions.

9. Report actioned content and proactively actioned content as a proportion of estimated violating content.

10. Break out actioned content measures by type of action taken.

11. Explore ways of accounting for changes in the Standards and changes in technology when reporting metrics in CSER.

12. Explore ways to enhance bottom-up (as opposed to top-down) governance, such as:

    - Create an elected "parliament" of users with powers over the definition of Community Standards.

- Create a broad partnership with a non-governmental organization specialized in the defense of democratic values and individual liberties.[25]
- Promote the creation of "nested communities" (building on the example of "groups") within which mechanisms of "bottom-up" governance tend to be more effective.[26]
- Organize regular third-party evaluations such as DTAG.

13 Enhance components of procedural justice (participation and voice, fairness, conveying trustworthy motives, treating people with respect and dignity) in the Community Standards enforcement and appeal/review process by taking measures such as:

- Providing more thorough explanations for decisions that a post violates, and for decisions that a post does not violate.
- Making appeal/review available not only for users whose posts are taken down, but also for users who flag a post as violating, and Facebook decides to leave the post up.
- Providing an opportunity to participate in the appeals/review process.
- Recruiting random panels of users or "citizens juries" to evaluate appeals.
- Making efforts to familiarize users with the Community Standards up front.
- Announcing forthcoming changes to the Community Standards in advance, and inviting users to weigh in on these changes.
- Finding ways to communicate more information to users about how content reviewers are trained and managed.

14 Publicly release anonymized or otherwise aggregated versions of the data that it uses to calculate prevalence and other metrics.

15 Modify the formatting, presentation, and text of CSER documents to make them more accessible and intelligible to readers by taking the following steps:

- Create permanent PDF versions of the different versions of CSER and make them easy to find on Facebook's site.
- Make it easier to locate other information related to Community Standards, such as the minutes from the Community Standards Forums, and announcements about future forums.
- Include more visual and narrative examples of challenging content-moderation decisions.
- Utilize interactive data displays, such as interactive maps or dashboards, as opposed to spreadsheets, statistics, and graphs.

[25] There are some examples of successful partnerships between NGOs and multinational companies, see e.g. the partnership between Chevron and WWF in the 1990s with positive environmental impacts over the Kikori forest in Papu New Guinea.

[26] Nobel Prize Recipient Elinor Ostrom (Economics) advocated the use of « nested enterprises » in order to better regulate large communities.