Oversight Board Comment, Case 2022-002-FB-MR

Mnemonic is the umbrella organization for the Yemeni Archive, Syrian Archive, and Sudanese Archive.[1] Our comments on this case are based on 11 years of experience in open source investigations, tracking of removal of content from social media platforms, and specific experience in Sudan. The Sudanese Archive came out of a collaboration with Sudanese organization Gisa.[2]

Facebook is the largest storehouse of open source documentation of human rights abuses in Sudan. Deletion of documentation is a constant threat.[3] The Board asks about Meta's policy on graphic and violent content. Meta users seldom know why their content was taken down, but the impacts are the same—permanent loss of potentially irreplaceable documentation. We see a significant amount removed even as we rush to archive and verify it for specific investigations such as our "June 3 Security Database."[4] As we explain, "By examining videos and photos from that day, Sudanese Archive documented more than 40 incidents in over twenty locations across and near the sit-in, and imagery of the use of batons and sticks, as well as the presence of rifles and anti-aircraft weapons."

We ask the Oversight Board to ensure that Meta's moderation policies and practices don't lead to loss of such essential documentation. Even when it is not directly admissible in court, open source human rights documentation can help investigators know where to start and what to look for. This content could aid the international justice mechanisms and domestic prosecutors. In fact, it has already been used in one domestic case to hold a perpetrator accountable. In August 2021 Judge Ahmed Hassan al-Rahma convicted six members of the country's paramilitary forces of the killing of 6 protestors in July 2019.[5] We were able to provide open source evidence in that case. Open source investigations could also help build cases for sanctions against specific perpetrators. Sudanese Archive is training lawyers and legal practitioners to be ready to present this evidence in cases in the future, with an eye to admissibility.

When we talk about open source investigations in Sudan, we cannot emphasize enough that every piece of content matters. In some places where we work, documentation is meticulously created by attorneys or other experienced witnesses. By contrast, in Sudan footage often comes from people in the midst of chaotic situations. Investigations thus rely on combining many

---

[1] Mnemonic, *About us*, https://mnemonic.org/en/about *last accessed* 29 Mar. 2022.
[2] Sudanese Archive, *About us*, https://sudanesearchive.org/about *last accessed* 29 Mar. 2022.
[3] Our archive includes many verified videos. We host some content directly on our own platforms after archiving and verification, but when looking at our database you can see videos that were previously on Meta platforms that have been removed. For example, incident "J3000146" in our database, labeled, "A video of RSF soldiers chasing and intimidating civilians" is no longer available on Facebook. (https://sudanesearchive.org/data/incidents/J3000146)
[4] Sudanese Archive, *June 3 Security Database,* https://sudanesearchive.org/datasets/june-03-security *last accessed* 29 Mar. 2022
[5] Noha Elhennaway, *Sudanese court sentences 6 to death for killing protesters*", AP News, Aug. 5, 2021, https://apnews.com/article/middle-east-africa-sudan-415b79e84cfaffcacf51f439cffd6d0f

"puzzle pieces" to form a complete picture. Livestreams are particularly important and helpful, as they are much easier to verify for authenticity, but even a well-shot livestream will need supporting documentation. For example, we may review a livestream that goes for 30 minutes where people are shot, but from the angle or chaos of the livestream it's not immediately apparent what happened. When we combine the livestream with later videos from a hospital, we can pair things like clothes and faces to show what happened. Finally, another important feature of footage from Sudan is that it is particularly likely to be graphic because of the nature of the settings and who is creating content. Journalists, for example, would employ tactics like shifting a camera to avoid details of a gruesome wound. Since most of the content we rely on is not created in this way, there is a greater risk of removal of some essential puzzle piece.

As a preliminary matter, currently Meta's "newsworthiness" exception is not written in a way that clearly covers graphic human rights content. Meta should specifically include human rights documentation in the policy. We believe this would lead to less removal of graphic content that is, nonetheless, important documentation. We do want to note that in Arabic-speaking countries like Sudan, Meta's "Dangerous Individuals and Organizations" (DIO) policy, which has been rightfully criticized by the Oversight Board in the past, also contributes to over-removal. According to Meta's own Community Standards, only clear "praise, support, and representation" should be removed. That does not mean that any mention of an organization or individual on the list is grounds for removal. In fact, as made clear by the company itself in response to Oversight Board case 2021-006-IG-UA, political discussion that is not praise, support, or representation about banned individuals and organizations is allowed under the policy. Meta should ensure that it is truly following its own policies on DIO and not taking down content that incidentally mentions groups or names on the DIO list. As we have documented over many years, DIO enforcement, especially when done by automated means, is a major threat to human rights documentation.[6]

In addition to improving and clarifying its policies, Meta must also greatly improve its content moderation processes in Sudan, and other places with ongoing conflicts. As the Facebook Papers confirmed, Meta tolerates an incredibly high rate of failure in the Arabic speaking world. As reported by Politico and confirmed by our own review of the Facebook Papers, "clunky algorithms to detect terrorist content incorrectly deleted non-violent Arabic content 77 percent of the time" while "[o]nly six percent of Arabic-language hate content was detected on Instagram before it made its way onto the photo-sharing platform owned by Facebook."[7] Sudan is no exception. In fact, Sudan appears to be one of the countries that has no direct dialect support. Meta must address insufficient support for Sudanese Arabic, discontinue or vastly improve weak

[6] Abdul Rahman Al Jaloud et al, *Caught in the Net: The Impact of "Extremist" Speech Regulations on Human Rights Content*, EFF, Syrian Archive, and WITNESS, 30 May 2019, *available online at* https://mnemonic.org/en/content-moderation/impact-extremist-human-rights

[7] Mark Scott, *Facebook did little to moderate posts in the world's most violent countries*, Politico, 25 Oct. 2021, https://www.politico.com/news/2021/10/25/facebook-moderate-posts-violent-countries-517050;

machine learning processes, and address targeted reporting from government and other bad actors.

Screenshots from a leaked Facebook document titled "Arabic dialect representation in markets":

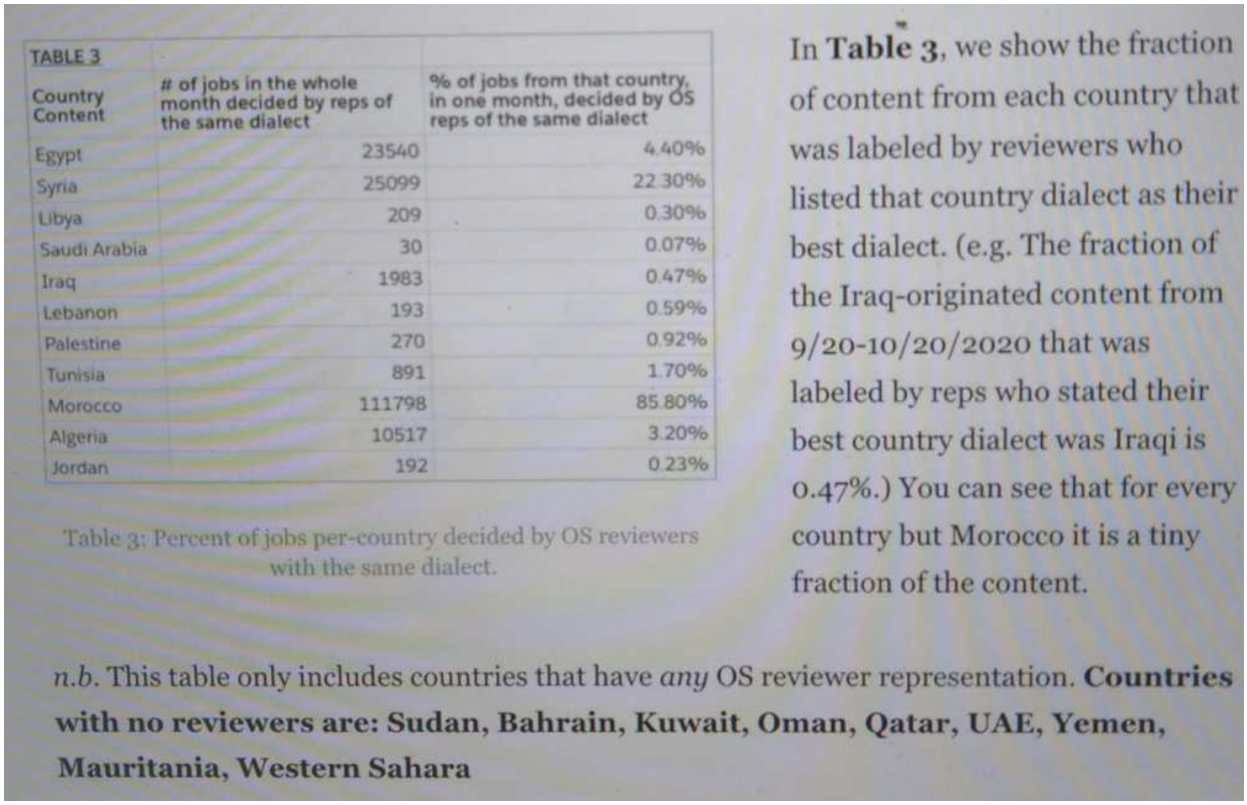


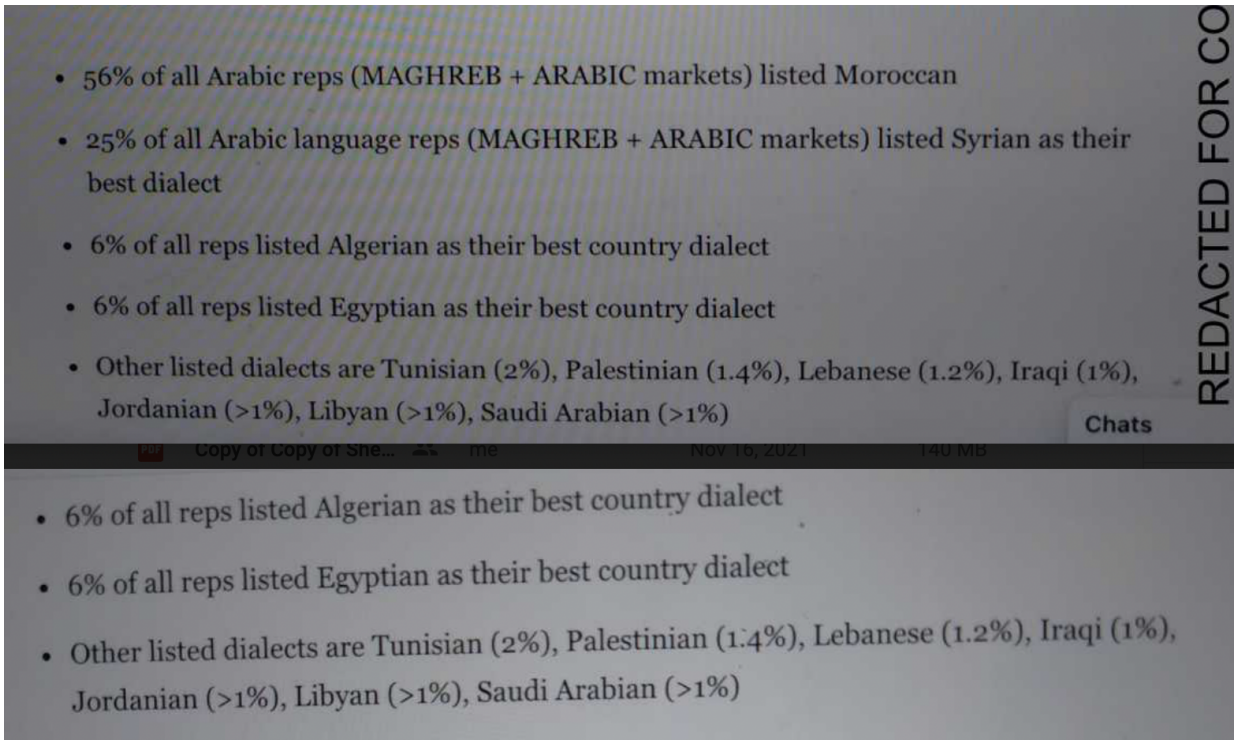
TABLE 3

Country Content	# of jobs in the whole month decided by reps of the same dialect	% of jobs from that country, in one month, decided by OS reps of the same dialect
Egypt	23540	4.40%
Syria	25099	22.30%
Libya	209	0.30%
Saudi Arabia	30	0.07%
Iraq	1983	0.47%
Lebanon	193	0.59%
Palestine	270	0.92%
Tunisia	891	1.70%
Morocco	111798	85.80%
Algeria	10517	3.20%
Jordan	192	0.23%

Table 3: Percent of jobs per-country decided by OS reviewers with the same dialect.

In **Table 3**, we show the fraction of content from each country that was labeled by reviewers who listed that country dialect as their best dialect. (e.g. The fraction of the Iraq-originated content from 9/20-10/20/2020 that was labeled by reps who stated their best country dialect was Iraqi is 0.47%.) You can see that for every country but Morocco it is a tiny fraction of the content.

n.b. This table only includes countries that have *any* OS reviewer representation. **Countries with no reviewers are: Sudan, Bahrain, Kuwait, Oman, Qatar, UAE, Yemen, Mauritania, Western Sahara**



- 56% of all Arabic reps (MAGHREB + ARABIC markets) listed Moroccan
- 25% of all Arabic language reps (MAGHREB + ARABIC markets) listed Syrian as their best dialect
- 6% of all reps listed Algerian as their best country dialect
- 6% of all reps listed Egyptian as their best country dialect
- Other listed dialects are Tunisian (2%), Palestinian (1.4%), Lebanese (1.2%), Iraqi (1%), Jordanian (>1%), Libyan (>1%), Saudi Arabian (>1%)

REDACTED FOR CO

Chats

Copy of Copy of She... me Nov 16, 2021 140 MB

- 6% of all reps listed Algerian as their best country dialect
- 6% of all reps listed Egyptian as their best country dialect
- Other listed dialects are Tunisian (2%), Palestinian (1.4%), Lebanese (1.2%), Iraqi (1%), Jordanian (>1%), Libyan (>1%), Saudi Arabian (>1%)

Meta can start to address its failures with moderating Arabic-language content by hiring more Sudanese dialect experts and using less machine learning. The Facebook Papers demonstrated how little support there is specifically for Sudanese dialects of Arabic (even within Sudan there are regional differences.) Bad moderation, especially when combined with a poor understanding of dialects, is compounded by the use of automation. Without crystal clear, high quality training data, problems are "baked in" to machine learning processes, leading to further improper takedowns. Meta must put in the work to identify and engage with civil society and language experts that can help address these problems, but it can't rely on free labor from civil society for this. It needs to spend more money to ensure that there is proper language support and that machine learning processes are of the highest quality.

Meta should also provide more protection from the targeted reporting rained on accounts that post documentation. This reporting can lead to removal of specific pieces of content or, even worse, entire accounts. This compounds verification problems- livestreams are much easier to verify and provide higher quality evidence, but users are forced to post content after the fact or on different platforms when their accounts are suspended. Meta has already demonstrated that it can provide protection to specific accounts through its Xcheck system.[8] It should provide a targeted layer of additional review to accounts during moments of unrest in Sudan.

The Board asks about Meta's use of age-gating and graphic violence warnings (aka interstitials). We often see interstitials applied in nonsensical and unfair ways. For example, content from protests that would be left up unobscured in the US, or even completely nonviolent content, may be put behind a warning in Sudan. We ask the Board to call on Meta to assess its' global application of interstitials to ensure that the policy is being applied in a fair and consistent way.

That being said, we note that there is not a consensus amongst activists about interstitials. Some feel that they make it more difficult to raise awareness and impinge on freedom of expression. Others agree that since this content can be traumatic, interstitials are a reasonable solution. As a group of people who sift through graphic content day in and day out, we are well aware of the reality of vicarious PTSD, as well as the fact that graphic content can raise awareness when human rights abuses are ignored or misunderstood by mainstream media. For the sake of preserving evidence while protecting mental health, we believe there are times when such interstitials are appropriate.

[8] "The cross-check system was built to prevent potential over-enforcement mistakes and to double-check cases where, for example, a decision could require more understanding or there could be a higher risk for a mistake. This could include activists raising awareness of instances of violence." Nick Clegg, *Requesting Oversight Board Guidance on Our Cross-Check System,* Facebook Newsroom, 28 Sep 2021, https://about.fb.com/news/2021/09/requesting-oversight-board-guidance-cross-check-system/

We also urge Meta to address privacy and dignity concerns in other ways. Often, people in Sudan and in other conflict zones are willing to have their image shared publicly in order to raise awareness about human rights violations. Meta can't know how people feel about this without asking them—and it can do so by providing an actual option for reporting images of oneself. This option should only be available for ones' own images, otherwise it would certainly be misused. Currently, "privacy" or "safety" are not available in either the Facebook or the Instagram reporting flow. That means that if someone clicks "report", they have to select "Something else," and would have to be familiar with the Community Standards.

Meta should make it clearer how to report images that violate its policy on "Privacy violations and Image Privacy Rights." Furthermore, the policy should be connected to Meta's "outing" policies that are meant to protect defectors and members of outing-risk groups.[9] These policies should be refined to consider the risks inherent to victims and witnesses of human rights violations. Finally, as it has done in Ukraine, Meta should increase its efforts to provide information and tools to users in Sudan and other places with ongoing protests, coups, and war about safety risks to themselves and others when documenting social movements and human rights abuses.[10]

[9] The "Outing" section of the "Coordinating Harm and Promoting Crime" policy prohibits "Content that exposes the identity or locations affiliated with any individual who is alleged to:Be a member of an outing-risk group; and/orShare familial and/or romantic relationships with a member(s) of an outing-risk group; and/or Have performed professional activities in support of an outing-risk group (except for political figures). Facebook Community Standards, Coordinating Harm and Promoting Crime, *last updated* 24 Feb. 2022, https://transparency.fb.com/policies/community-standards/coordinating-harm-publicizing-crime//. The "Privacy Violations" policy covers specific types of information, including images of minors. It also covers "Content that puts a defector at risk by outing the individual with personally identifiable information when the content is reported by credible government channels" and "Depictions of someone in a medical or health facility if reported by the person pictured or an authorized representative." Facebook Community Standards, Privacy Violations, *last updated 24 Feb. 2022,* https://transparency.fb.com/policies/community-standards/privacy-violations-image-privacy-rights/

[10] Nathaniel Gleicher and David Agranovich, *Updates on Our Security Work in Ukraine*, Facebook, *last updated* 27 Feb. 2022, https://about.fb.com/news/2022/02/security-updates-ukraine/