HARVARD Kennedy School
CARR CENTER
for Human Rights Policy

SPRING 2024
ISSUE 2024-01

CARR CENTER FOR HUMAN RIGHTS POLICY
HARVARD KENNEDY SCHOOL

Does AI Understand Arabic?

Evaluating the Politics Behind the Algorithmic Arabic Content Moderation

Mona Elswah

Carr Center
Discussion Paper

Does AI Understand Arabic?

Evaluating the Politics Behind the Algorithmic Arabic Content Moderation

Carr Center for Human Rights Policy
Harvard Kennedy School, Harvard University
January 30, 2024 | Issue 2024-01

Mona Elswah
Technology and Human Rights Fellow, Carr Center for Human Rights Policy

ABSTRACT

Artificial Intelligence (AI) continues to be integrated into various domains and industries. Over the years, social media companies have utilized AI technologies to moderate users' content, personalize recommendations, and optimize overall user experience. While machine learning models have been found effective in identifying and addressing harmful and violent content, a mounting number of concerns were raised regarding the bias and discriminatory decisions made by these models when applied to non-English content. In this paper, I zoom in on the AI-powered content moderation by Meta's Facebook in relation to managing Arabic content. I argue that the Arabic content is subject to "inconsistent moderation," meaning that some content will be over-moderated, while other content will be left untouched despite violating the platforms' standards. These inconsistencies have limited users' ability to engage in meaningful political debates in the region. Put simply, Arabic-speaking users are now uncertain whether their content will be deleted or kept by the algorithm. This type of unclear and inconsistent moderation has led to a social distrust towards AI tools and applications among Arab Internet users.

Above: Arabic inscriptions at the Bloomingdale's on Wabash (formerly the Medinah Center) in Chicago. Image credit Atramos.

> **"We believe that all people are equal in dignity and rights."**
> — Meta, 2023

I. Introduction

AI-powered (also known as, "algorithmic" or "automated") content moderation is crucial to helping platforms manage content in a short time window at a fraction of the cost.[1] This system depends on machine learning models to classify users' content and to make decisions about it. Yet, many activists and civil society leaders complained about the alarming number of errors caused by algorithmic content moderation, accusing Silicon Valley companies of censorship, bias, and disillusionment.[2] In several instances, users reported the "over-moderation" of their content whereby the algorithm removed or suspended their accounts after they promoted political issues and opposed their local governments.[3] However, social media companies have claimed that these were merely "technical errors" and were not intentionally gatekeeping or censoring political discourse in the Arab world.[4] On the other hand, reports showed that there are numerous "language blind spots" on Facebook that have made some of the Arabic content "under-moderated." These blind spots have allowed hate speech, conspiracies, and terrorist content to evade moderation efforts.[5]

Despite the considerable attention brought to the errors and discrimination of AI-powered content moderation, only a few studies have examined the platforms' algorithmic design in the Arabic-speaking world. Arabic is among the most frequently used languages on the Internet and the sixth most spoken language in the world. However, tech companies are paying less attention to this language while developing their policies and machine learning tools.

In this article, I zero in on Facebook's Arabic algorithmic content moderation, aiming to understand when does AI take down content, and whether Facebook over-moderates or under-moderates Arabic. I only focus on Facebook—a platform managed and owned by Meta—because a) it is the most impactful social media platform in the Arab world, with far more complaints and controversy; b) it is considered the most restrictive in terms of moderating Arabic content;[6] and c) Meta's products employ the same algorithms to police Arabic language online across its different products, according to the interviews I conducted with digital rights advocates and page administrators (page admins). In this article, I focus on political content, being the most controversial and subjected to a significant number of takedowns and restrictions in the last five years.

I assess the Arabic content administered by Facebook using two methods. I first analyzed Facebook's official policies, BSR[7] and Oversight board reports, leaked internal documents, and civil society groups' reports. In addition, I met several civil society members who are devoting their time to fighting Facebook's bias and Arab Facebook pages admins who were impacted by these policies.

---

[1] Robert Gorwa, Reuben Binns, and Christian Katzenbach, "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance," *Big Data & Society* 7, no. 1 (January 1, 2020): 2053951719897945, https://doi.org/10.1177/2053951719897945.

[2] Ahmed Qadi, Mona Shtaya, and Cathrine Abuamsha, "Hashtag Palestine 2022: 7amleh Restored a Third of the Deleted Palestinian Content from Social Media Platforms," *7amleh* (2023), https://7amleh.org/2023/02/02/hashtag-palestine-1119-palestinian-digital-rights-violations-during-the-year-2022.

[3] Ryan Mac, "Al-Aqsa: Instagram Labeled One of Islam's Holiest Mosques a Terrorist Org," *Buzzfeed* (2021), https://www.buzzfeednews.com/article/ryanmac/instagram-facebook-censored-al-aqsa-mosque; Jillian C. York, *Silicon Values: The Future of Free Speech Under Surveillance Capitalism* (New York: Verso Books, 2022).

[4] Access Now, "Open Letter to Facebook on the Upcoming Tunisian Elections of 2019," *Access Now* (blog) (September 2, 2019), https://www.accessnow.org/open-letter-to-facebook-regarding-the-upcoming-tunisian-elections-of-2019/.

[5] Moustafa Ayad, Anisa Harrasy, Mohammed Abdullah A., "Under-Moderated, Unhinged and Ubiquitous: Al-Shabaab and the Islamic State Networks on Facebook," ISD (June 14, 2022), https://www.isdglobal.org/isd-publications/under-moderated-unhinged-and-ubiquitous-al-shabaab-and-the-islamic-state-networks-on-facebook/.

[6] Qadi, Shtaya, and Abuamsha, "Hashtag Palestine 2022."

[7] Dunstan Allison-Hope, Jenny Vaughan, and Lindsey Andersen, "Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021" BSR (2022), https://www.bsr.org/en/blog/human-rights-due-diligence-of-meta-impacts-in-israel-and-palestine-may-2021.

"Arabic has a rich morphology, meaning that a single word can have several meanings and variations that change depending on the context."

After many interviews, observations, and document analysis, I argue that Facebook employs "inconsistent moderation" in relation to Arabic-language content. This inconsistency has reinforced a form of social distrust towards AI tools and applications among Arab Internet users. This social distrust in the AI system will take years to overcome and can impact all AI-powered tools. This article contributes to the discussion on platform governance and AI accountability, highlighting a region that is under-explored in academic scholarship.

II. Does AI Understand Arabic?

Arab AI users face profound language discrimination while dealing with AI tools that are mainly available in English or provide higher quality outputs for English-speaking users. This language gap restricts the creativity and productivity of Arab AI-users, depriving them from the same opportunities as their counterparts in the English-speaking world. For example, AI-powered video making apps do not regulate the direction of the Arabic language—from right to left—and reverse the position of the words thereby rendering them incomprehensible.[8]

There are many reasons why it is challenging to build AI models that could produce accurate AI-powered services to the MENA region. Arabic is an expressive language that derives its roots from a diverse set of languages, and exhibits complex morphologies, phonologies, and semantic and syntactic structures.[9,10] It is a language that is spoken by a large population of more than 400 individuals with dialectal variations.[11]

Arabic has a rich morphology, meaning that a single word can have several meanings and variations that change depending on the context. Unlike English, Arabic has also a complex syntax, allowing a flexible word order which makes it challenging for AI models to understand the relationship between words in a sentence.[12] In addition, the enormous dialectal variations, and limited resources and research conducted to examine Arabic, pose extra challenges to process Arabic language.[13]

The recent advancement in Large Language Models (LLM), led by ChatGPT and Bard, allowed the generation of multilingual human-like responses.[14] Despite the admirable performance of the OpenAI's ChatGPT in dealing with English-language content, researchers found that ChatGPT 3.5 is still behind in dealing with multi-dialectal Arabic content. Researchers found that smaller dedicated models for Arabic did a better

"Researchers found that smaller dedicated models for Arabic did a better job than ChatGPT.[15] This indicates that the problem does not lie in the inherent difficulty of Arabic but rather in the level of dedication and willingness to invest in improving AI models that meet Arabic's unique characteristics."

[8] Amal Mekki, "The AI Revolution Is Leaving Arabic Speakers Behind," *Middle East Eye* (2023), https://www.middleeasteye.net/opinion/al-revolution-arabic-speakers-left-behind.

[9] Md Tawkat Islam Khondaker et al., "GPTAraEval: A Comprehensive Evaluation of ChatGPT on Arabic NLP" arXiv (May 24, 2023), http://arxiv.org/abs/2305.14976.

[10] Ali Alshehri, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed, "Understanding and Detecting Dangerous Speech in Social Media," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, with a Shared Task on Offensive Language Detection (OSACT 2020, Marseille, France: European Language Resource Association, 2020), 40–47, https://aclanthology.org/2020.osact-1.6.

[11] Mahmoud El-Haj, "Habibi - A Multi Dialect Multi National Arabic Song Lyrics Corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference* (LREC 2020, Marseille, France: European Language Resources Association, 2020), 1318–26, https://aclanthology.org/2020.lrec-1.165.

[12] Sahar Al Muscati, "The State of AI in the Arab World," *WAYA* (blog), (April 16, 2023), https://waya.media/the-state-of-ai-in-the-arab-world/.

[13] Al Muscati.

[14] Khondaker et al., "GPTAraEval."

job than ChatGPT.[15] This indicates that the problem does not lie in the inherent difficulty of Arabic but rather in the level of dedication and willingness to invest in improving AI models that meet Arabic's unique characteristics. Furthermore, it is worth noting that Google's Bard did not support Arabic in its first version until its update in July 2023. Arabic was later introduced in their update, indicating that its inclusion was a "nice addition" by the company rather than a mandatory feature in the earlier release of the AI tool.

Arabic was never a top priority to AI developers and is unlikely to become one in the near future. Consequently, this might impede the creativity and innovation of Arab Internet users. Moreover, it may result in many mistakes and errors. On the one hand, it could lead to censoring and restricting Arab users' freedom of expression, with their social media posts being mistakenly flagged and removed by the algorithms. On the other hand, the poor design of AI tools could also lead to the spread of misinformation and hate speech, leaving such content without removal.

III. AI-Powered Arabic Content Moderation

Academic scholarship that once credited social media for enabling democracy in the region has shifted to blaming online networks for censoring and restricting Arabic content. The architectural design and technical affordances of social media platforms were criticized for empowering governments and obstructing activists and civil society groups.[16] The failures and subjectivity of social media platforms have led to the rise of digital orientalism by disadvantaging users in the Arab region.[17] Recently, automated content moderation has become the subject of interest after its inherent inequity in regulating misleading Arabic content,[18] removing war evidence content,[19] and suppressing human rights defenders.[20]

In my past work, I found that Facebook and Instagram have systematically and intentionally silenced the voices of activists,[21] designed algorithms that failed to fit the complexity of the Arabic language,[22] and neglected the regions' civil society

[15] Khondaker et al.

[16] Mona Elswah and Philip N. Howard, "The Challenges of Monitoring Social Media in the Arab World: The Case of the 2019 Tunisian Elections," The Computational Propaganda Project (2020), https://comprop.oii.ox.ac.uk/research/tunisia-election-memo/; Olivia Solon, "How Syria's White Helmets Became Victims of an Online Propaganda Machine," *The Guardian* (December 18, 2017), https://www.theguardian.com/world/2017/dec/18/syria-white-helmets-conspiracy-theories.

[17] Mahsa Alimardani and Mona Elswah, "Digital Orientalism: #SaveSheikhJarrah and Arabic Content Moderation," *POMEPS Studies 43: Digital Activism and Authoritarian Adaptation in the Middle East* (August 5, 2021), https://papers.ssrn.com/abstract=3900520.

[18] Paul Mozur, "A Genocide Incited on Facebook, with Posts from Myanmar's Military," *The New York Times*, (October 15, 2018), https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.

[19] William Lafi Youmans and Jillian C. York, "Social Media and the Activist Toolkit: User Agreements, Corporate Interests, and the Information Infrastructure of Modern Social Movements," *Journal of Communication* 62, no. 2 (April 2012): 315–329, https://doi.org/10.1111/j.1460-2466.2012.01636.x

[20] York, *Silicon Values*.

[21] Alimardani and Elswah, "Digital Orientalism."

[22] MENA Region," *Social Media + Society 6*, no. 3 (July 1, 2020), https://doi.org/10.1177/2056305120948251.

"In many instances, Facebook's algorithmic moderation has censored and deleted activists' content. In June 2020, Facebook deleted dozens of accounts belonging to Syrian, Palestinian, and Tunisian activists and journalists who work in the human rights field. These accounts were miscategorized as being linked to terrorism by the algorithm and have not been retrieved to this date."[25]

—

Mona Elswah
Technology and Human Rights Fellow

[25] Olivia Solon, "'Facebook Doesn't Care': Activists Say Accounts Removed Despite Zuckerberg's Free-Speech Stance," NBC News (June 15, 2020), https://www.nbcnews.com/tech/tech-news/facebook-doesn-t-care-activists-say-accounts-removed-despite-zuckerberg-n1231110.

demands.[23] It was also reported that Facebook and Instagram had adjusted their algorithms to reduce activists' views and reach, undermining freedom of expression.[24]

In many instances, Facebook's algorithmic moderation has censored and deleted activists' content. In June 2020, Facebook deleted dozens of accounts belonging to Syrian, Palestinian, and Tunisian activists and journalists who work in the human rights field. These accounts were miscategorized as being linked to terrorism by the algorithm and have not been retrieved to this date.[25] In May 2021, hundreds of Palestinian activists struggled to get their voices heard on Instagram and Facebook while the algorithms removed the majority of their views that criticize Israel—a phenomenon that was referred to as "digital apartheid."[26]

While activists are pressuring tech companies to keep their content, conspiratorial and hate speech content has managed to circumvent AI moderation. Tactics have evolved to evade AI-powered moderation in order to spread divisive content, leading some to accuse Facebook and other platforms of under-moderating Arabic content. For instance, Arabic-language pages with terrorism-related content would identify themselves as "educational" to be able to spread terrorist materials, like books and information, without being removed. Educational pages are less likely to be taken down, unless found to be violating the community standards.[27]

To this date, we are uncertain about why and what content gets removed by AI and whether Facebook over-moderates or under-moderates the Arabic language. To answer these questions and more, I take a deep dive into Facebook's governance and its algorithmic design.

"To this date, we are uncertain about why and what content gets removed by AI and whether Facebook over-moderates or under-moderates the Arabic language."

IV. Inspecting Facebook Governance

A) MODERATING CONTENT

Technology companies do not publicly share their algorithmic moderation policies. Usually, it is hard to know how the algorithm was designed, what kind of training data was fed to the machine learning model (ML), and who is behind this design. While we often do not know the answers to these questions, technology companies share their community standards (CS)[28] and their policies to govern violence and incitement (V&I).[29] In addition, leaks and whiste-blowers allow policymakers to understand a little bit about the AI-powered moderation system that a company uses.

Hence, I decided to examine the documents released by Facebook in relation to their content moderation system, such as CS, V&I,[30] and the Oversight Board reports. I also relied on the leaked list of Facebook's dangerous organizations and

[23] Elswah and Howard, "The Challenges of Monitoring Social Media in the Arab World."

[24] Mac, "Al-Aqsa: Instagram Labeled One of Islam's Holiest Mosques A Terrorist Org."

[25] Olivia Solon, "'Facebook Doesn't Care': Activists Say Accounts Removed Despite Zuckerberg's Free-Speech Stance," *NBC News* (June 15, 2020), https://www.nbcnews.com/tech/tech-news/facebook-doesn-t-care-activists-say-accounts-removed-despite-zuckerberg-n1231110.

[26] 7amleh, "The Attacks on Palestinian Digital Rights," (May 2021), https://7amleh.org/storage/The%20Attacks%20on%20Palestinian%20Digital%20Rights.pdf.

[27] Moustafa Ayad, "Facebook and YouTube Are Failing to Detect Terrorist Content in Arabic," *Vice* (blog) (August 9, 2019), https://www.vice.com/en/article/59nmyd/facebook-and-youtube-are-failing-to-detect-terrorist-content-in-arabic.

[28] Facebook, "Facebook Community Standards," Meta (2023), https://transparency.fb.com/en-gb/policies/community-standards/.

[29] Transperancy Center, "Violence and Incitement," (2023), https://transparency.fb.com/en-gb/policies/community-standards/violence-incitement/.

[30] It is worth mentioning that the Localization Lab and Internews analyzed the Arabic versions of CS and V&I and found several translation and context errors. These errors included mistranslation of some words, punctuation, and providing Western-based examples in the text. The full analysis can be found here: Localization Lab and Internews, "'Wait, Who's Timothy McVeigh?': A Translation Review of Facebook and YouTube Content Moderation Policies in Amharic, Arabic, Bengali, and Hindi," (accessed September 14, 2023), https://www.localizationlab.org/tech-policy-review.

individuals (DOI)[31] and the assessment that was carried out by Business for Social Responsibility (BSR). I also analyzed the reports and documents written by the two most important civil society organizations that were on the frontlines in the fight against Arabic content moderation bias: 7amleh - The Arab Center for the Advancement of Social Media, and Social Media Exchange (SMEX).

According to Facebook, the platform is committed to authenticity, privacy, dignity, and safety. Their community standards are applied to all types of content across the world. Removal of content depends on many factors, including if the content: a) potentially leads to a genuine risk of physical harm; b) incites or facilitates violence; c) praises, supports, or represents (PSR) a terrorist organization; and d) represents a direct threat to public safety. A threat could be an intent of violence, a symbol that represents violence, or a clear statement advocating for or admitting to high-severity violence. However, all of this depends on the context in which the content was written:

> "We also try to consider the *language and context* in order to distinguish casual statements from content that constitutes a credible threat to public or personal safety. In determining whether a threat is credible, we may also consider additional information such as a person's public visibility and the risks to their physical safety." —*Facebook CS, 2023.*

One of the main signals for the AI model to remove certain content is when it has a PSR reference to one of the entities or individuals on Facebook's DOI list. The DOI list includes organizations and individuals that conduct or have ties to offline and online violence, according to US foreign policy. Going against the Oversight Board's recommendation, Facebook has repeatedly refused to publish this list, fearing it will endanger its employees. This 100-page-list—which was leaked in 2021[32]—has been criticized heavily for disproportionately punishing certain communities and organizations and undermining freedom of expression.[33] In many instances, it constrains political debates about some Arab countries' leadership. Although news and neutral statements are not subject to removal or suspension by the company, the content creators and digital rights activists I spoke with expressed the opposite.

B) METHODS OF MODERATION

To employ its policies, Facebook has incorporated four ways to moderate Arabic content: direct government requests,[34] algorithms, community reports, and fact checking (See Figure 1). Direct government requests to Facebook or Instagram are first reviewed to see if this content violated the community standards. If this was the case, the content will be removed from the platform globally and in all countries. If this is not the case, the content will be assessed by legal and human rights teams at Facebook to decide whether it violated a local law in the country where the user is based. The user, at this stage, can still appeal against this decision. Facebook and Instagram received more than 89,00 requests globally in 2022 alone. It has also been reported that Facebook has received governments' direct requests to delete content. This process is managed by human moderators and reviewers without intervention from the AI.

The second and most significant method to remove content is the machine learning model that Facebook developed to manage content. Facebook was admiring how their AI manages to remove more than 90% of the violating content before users report it.[35] In other words, their algorithm is the most important tool for managing content and is responsible for the majority

[31] Sam Biddle, "Revealed: Facebook's Secret Blacklist of 'Dangerous Individuals and Organizations,'" *The Intercept* (2021), https://theintercept.com/2021/10/12/facebook-secret-blacklist-dangerous/.

[32] "Facebook Dangerous Individuals and Organizations List," (2021), https://s3.documentcloud.org/documents/21083819/facebook-dangerous-individuals-and-organizations-list-reproduced-snapshot.pdf.

[33] Biddle, "Revealed."

[34] Facebook, "Content Restrictions Based on Local Law," Meta (2023), https://transparency.fb.com/data/content-restrictions/.

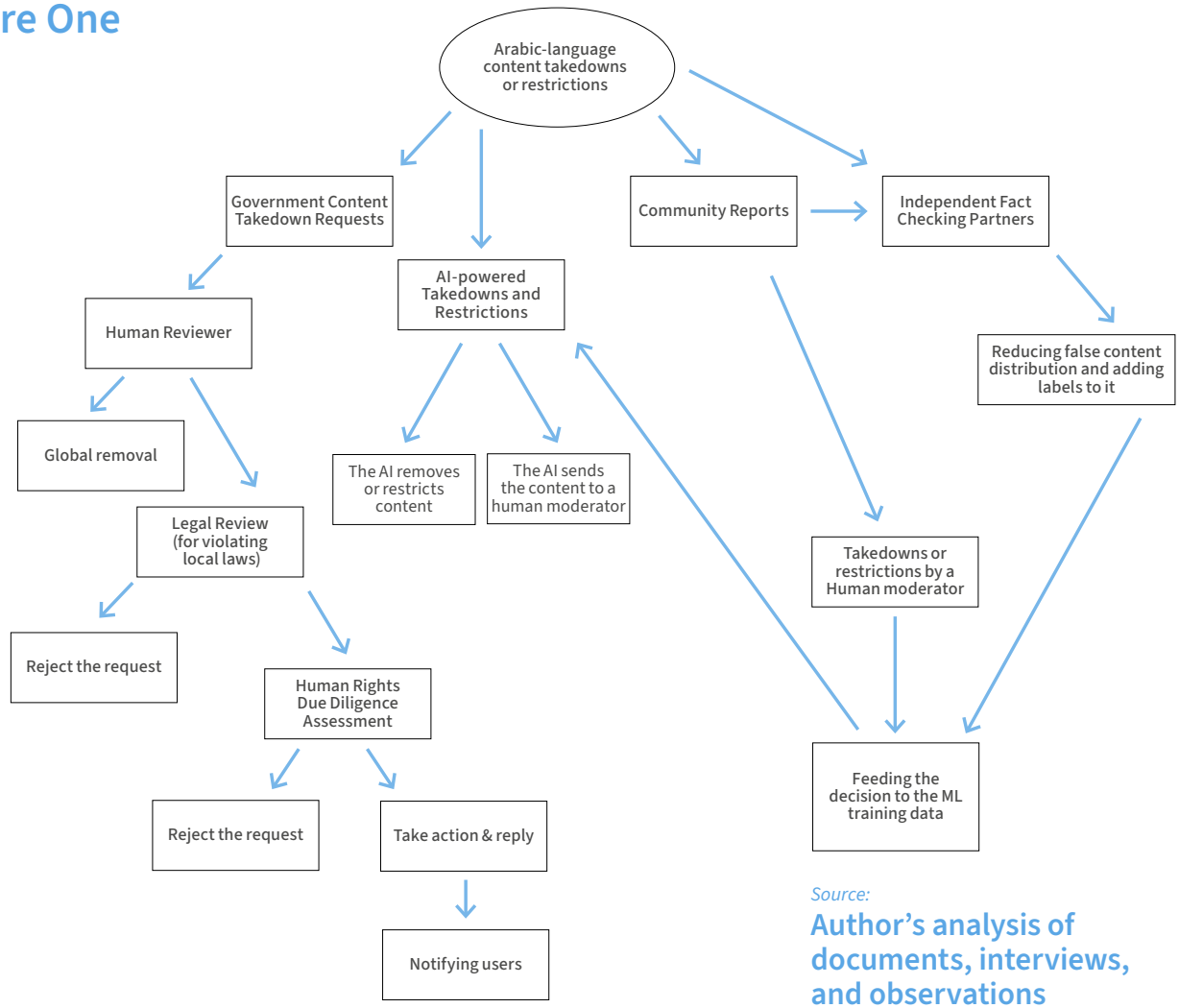
[35] Facebook, "How Technology Detects Violations," Meta (January 2022), https://transparency.fb.com/en-gb/enforcement/detecting-violations/technology-detects-violations/.

of the content restrictions and takedowns. When I inspected the system from the documents I read, I found that the AI plays three roles in content restrictions: a) direct removals of content based on the CS; b) reviewing some of the users' reports and making decisions about the content; and c) assigning the users' reports to human reviewers based on the severity of the content, when the AI fails to make a decision on its own.

The third method is the users' reports. In addition to machine learning-powered automated flagging, platforms depend on their user community to flag content. These reports are then checked by a human moderator to discern whether the reported content violates "community standards."[36] The significance of this method, according to the BSR report, is that it is the main source feeding the training data to improve the Arabic classifier. At many instances, the human moderators were found insufficient in making decisions about Arabic content from different dialects and contexts—especially the Palestinian dialect.[37] Those decisions would later feed the algorithm, causing the AI to make more wrong decisions.

The last method, which was incorporated in 2016, is partnering with fact checkers in the region. So far, Facebook has made partnerships with two fact checking organizations: AFP and Fatabyyano.[38] Fact checkers will either check for content using their own methods of investigation or Meta will send them content to classify as true or false. This content might have been flagged by a user or selected due to other predetermined

Figure One



Source:
Author's analysis of documents, interviews, and observations

Notes: This process has not been verified by Facebook and relies on the author's own analysis

[36] Abdul Rahman Al Jaloud et al., "Caught in the Net: The Impact of 'Extremist' Speech Regulations on Human Rights Content" Electronic Frontier Foundation, Syrian Archive, and Witness (May 2019), https://syrianarchive.org/assets/tech-advocacy/caught_in_the_net_whitepaper_2019.pdf.

[37] BSR (Business for Social Responsibility), "Human Rights Impact Assessment: Facebook in Myanmar," Facebook (October 2018), https://about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf.

[38] Facebook, "Where We Have Fact-Checking," Meta (accessed July 11, 2023), https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking/partner-map.

signals (e.g., comments and rapid distribution). If content was identified as false by a fact checker, Meta reduces its distribution and use this information to train their ML models.[39]

C) ALGORITHMIC MODERATION FLAWS

As mentioned earlier, Arabic is one of the world's most complicated languages. A name on the DOI list could be used for different reasons and in different contexts. The real problem that the civil society representatives I interviewed raised was how the AI model would not distinguish between what was a PSR to those on the DOI list and what is not (e.g., news). In a leaked document by the company's whistle-blower Frances Haugen, it was reported that the counter terrorism classifiers mistakenly deleted Arabic non-violent content 77% of the time.[40] In addition, about 48% of the disabled ads in the region were removed by the terrorism classifiers mistakenly.[41]

From the interviews, SMEX members told me that they would receive hundreds of complaints every year from Facebook users—many are journalists—whose content would be taken down for referencing one of the names on this DOI list while publishing a news story on their personal pages. These complaints showed that the AI-powered moderation disregarded the context of the content and did what respondents referred to as "over-moderation."

For example, a digital rights activist I interviewed told me about an incident when a happy new father posted a photo of his newborn "Qassem"—an Arabic name for a male—on his personal Facebook page. The photo was removed for being associated with "al-Qassam Brigades"—a military wing of the Palestinian organization Hamas. Any mention of the word "Hezb" (translated as "party") will be deleted for being mistakenly related to the Lebanese organization Hezbollah. Other incidents included the removal of any content related to the well-known Al-Aqsa mosque, the third holiest site in Islam, for being mistakenly associated with Al-Aqsa Martyrs' Brigades. This repeatedly restricted access to information about these entities and sharing their news on the region's most important platform.

While activists complained about the over-moderation of their content, reports showed that bad actors and terrorist organizations weaponized the platform's AI "blind spots." A leaked document showed that Facebook failed to protect users from terrorist and hate content that kept on spreading without detection.[42] Only 40% of Arabic hate speech content on Facebook was deleted by the hate speech classifiers,[43] leaving a plenty of hate content under-moderated and a space for terrorist organizations to evolve their strategies to combat the algorithmic moderation. For example, a report was published last year pointing out how Arabic pages affiliated with ISIS, posing as mainstream media pages, are using emojis instead of words to escape Facebook moderation.[44]

> "In addition to the inconsistency in moderating violent and non-violent content, the algorithmic moderation is also inconsistent across the region… moderation flaws varied across dialects."

In addition to the inconsistency in moderating violent and non-violent content, the algorithmic moderation is also inconsistent across the region. According to the interviews, the AI moderation is more likely to target the Mashriqi Arabic content (the dialect spoken in Arab countries located in the eastern part of the Arab world (e.g., Egypt, Iraq, Syria, and Saudi Arabia)). The counter terrorism classifiers over-mod-

[39] Facebook, "How Meta's Third-Party Fact-Checking Program Works," Meta (accessed July 11, 2023), https://www.facebook.com/facebookmedia.

[40] "Copy of Copy of MENA Integrity_sanitized - DocumentCloud," (2021), https://www.documentcloud.org/documents/21090828-copy-of-copy-of-mena-integrity_sanitized#document/p3/a2061347.

[41] "Copy of Copy of MENA Integrity_sanitized - DocumentCloud."

[42] Mark Scott, "Facebook Did Little to Moderate Posts in the World's Most Violent Countries," Politico (October 25, 2021), https://www.politico.com/news/2021/10/25/facebook-moderate-posts-violent-countries-517050.

[43] "Copy of Copy of MENA Integrity_sanitized - DocumentCloud."

[44] Mark Scott, "Islamic State Evolves 'Emoji' Tactics to Peddle Propaganda Online," Politico (February 10, 2022), https://www.politico.eu/article/islamic-state-disinformation-social-media/.

erate Syrian and Palestinian Arabic content[45] indicating that the AI model's training data excluded the Maghrebi Arabic (the dialect spoken in Arab countries located in the western part of the Arab world (e.g., Tunisia, Morocco, and Algeria)). In other words, moderation flaws varied across dialects.

AI-powered moderation also varied according to the source. From the interviews, AI over-moderates new and unverified Arabic-language pages and accounts. Content posted by verified pages with a large number of followers is less likely to be removed by the AI. However, if the same content is posted by a relatively new page with a smaller number of followers, it might be regulated or taken down entirely by the algorithm. For instance, one of the respondents, a moderator for many Arabic news Facebook pages, noted that when she posted a news story on a page she manages about the assassination of Japanese Prime Minster Shinzo Abe—the same story was shared by SkyNews Arabic—the page quality tab showed that the page became at risk and its reach was reduced drastically. The admin's own personal account was also sanctioned and she was not allowed to comment or post on Facebook for an entire month.

Conclusion: Looking Ahead

Does AI understand Arabic? The answer is yes, it can. However, the extent to which it comprehends Arabic depends on how well-trained the AI tool is. To-date, tech companies tend to under-train their AI tools to comprehend this language, impacting the accuracy and efficacy of these tools in dealing with the Arabic language. Despite that, tech companies are still heavily relying on AI-powered content moderation to regulate users' content.

In this paper, I reviewed documents that were leaked from Facebook as well as documents that Facebook published, and I spoke with civil society groups and page admins. I found that Facebook employs "inconsistent moderation." This inconsistency was evident in relation to the dialect, source, and context. Facebook performs over-moderation and under-moderation at the same time. Facebook over-moderated and mistakenly removed activists' content in many instances. At the same time, Facebook unintentionally under-moderated hate and terror content on various occasions. This over and under moderation were both caused by the poorly designed AI classifiers. They were both a result of mistakes and errors by the AI and the absence of sufficient motivation to improve the models and training data. Consequently, bad actors took advantage of this and were able to evolve their strategies online to evade the AI moderation and spread hate content on the platform.

Google and Facebook headquarters in Mountainview and Menlo Park. Image credit David Nagle and Minette Lonsie.

[45] "Copy of Copy of MENA Integrity_sanitized - DocumentCloud."

This paper also concludes the following:

- *The DOI list is overrepresenting Arabs and Muslims. This overrepresentation can lead to restrictions on users' experiences and the over-moderation of Arabic content.*
- *One of the major problems that Meta is having is the way it outsources its human moderators. Those moderators are the core source of the training data for the Arabic classifiers. They need better training on how to assess content from the different countries and dialects in the region.*
- *The lack of transparency is another problem that Facebook needs to fix urgently. The Oversight Board has repeatedly called for publishing the DOI list. However, Facebook has ignored that recommendation. Civil society groups have repeatedly demanded that Facebook reveal the algorithm and how it works. That has not been shared publicly as well.*
- *There is no easy way to fix this. Meta, the company that runs Facebook and many other platforms, is willing to invest in new platforms, like Threads, but claims that it does not have the capacity to improve their algorithmic content moderation for Arabic content. It is not a problem of the uniqueness of the Arabic language, but a problem of discrimination and bias imposed by tech companies against non-English content.*

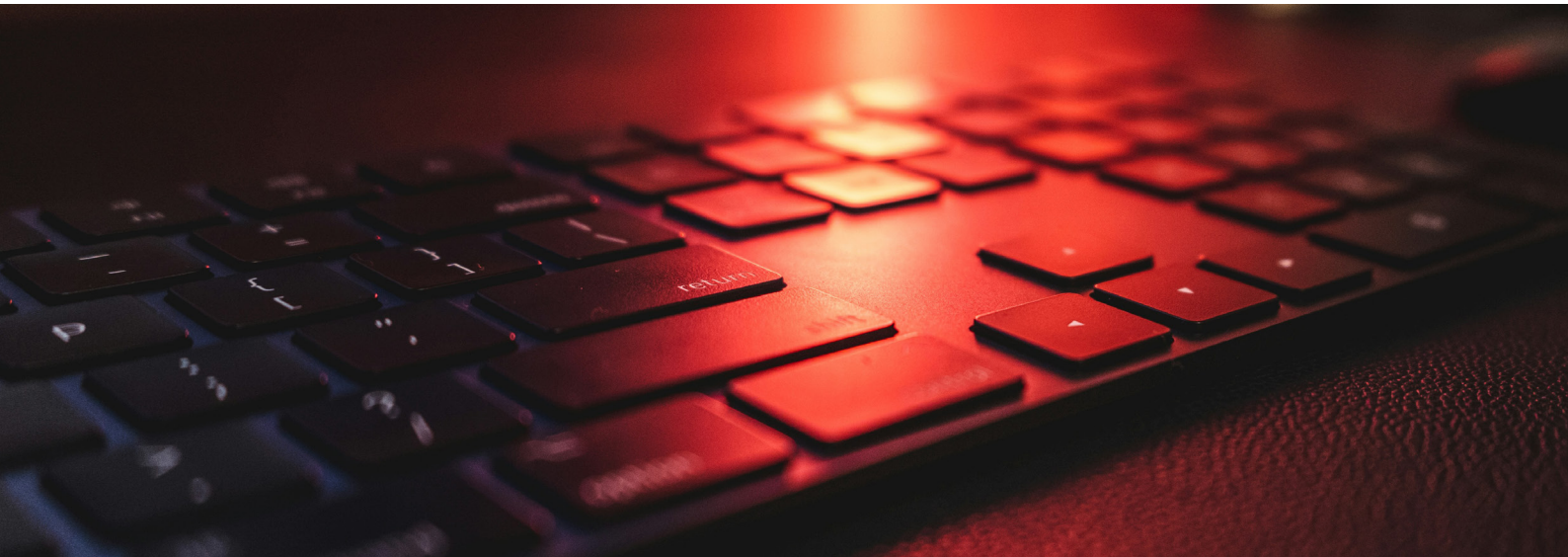
Based on the interviews and observations, I recommend the following:

A) While tech companies do consult a few civil society groups and other stakeholders in the region, members of these groups feel frustrated because they believe that their recommendations are usually ignored and not employed. This practice is soon going to discourage engagement and stakeholders will refuse to contribute. "Closing the feedback loop"46 is necessary to inform civil society groups and other consulted shareholders with the decisions that were made based on their feedback.

B) While consulting and engaging with civil society groups in the region is a good practice, it is undoubtedly crucial to increase recruitment of experts from the region who are proficient in both the dialectal and Modern Standard Arabic. By bringing in more engineers and data scientists with relevant linguistic and regional experience, the performance of the AI models can be significantly enhanced.

C) There is a pressing need for a more efficient and expedient appeals mechanism for users. Currently, social media users whose content is erroneously removed by the algorithms often find themselves facing a complicated and a protracted appeal process, which proves to be inefficient in many cases. As a result, users often turn to civil society groups to connect them with tech companies to retrieve their content, as clear and efficient methods for appealing decisions are lacking. These appeals must be given due consideration to improve the algorithms to prevent both over- and under-moderation of content. ■

[46] UN B-Tech Project, "Five Practices to Improve Stakeholder Engagement in Tech Company Due Diligence," UN Office of the High Commissioner for Human Rights (2022), https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/btech-stakeholder-engagement-paper.pdf.



Carr Center for Human Rights Policy
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

This publication was published by the Carr Center for Human Rights Policy at the John F. Kennedy School of Government at Harvard University

carrcenter.hks.harvard.edu

79 JFK Street | Cambridge, MA 02138
carrcenter@hks.harvard.edu