THE PROMISE INSTITUTE
FOR HUMAN RIGHTS

SOCIAL MEDIA, CONTENT MODERATION AND INTERNATIONAL HUMAN RIGHTS LAW:

THE EXAMPLE OF THE NAGORNO-KARABAKH/ARTSAKH CONFLICT

Aya Dardari, Nicholas Levsen, Ani Setian and Jessica Peake

ABOUT THE PROMISE INSTITUTE

The Promise Institute for Human Rights at UCLA School of Law is the center of human rights education, research and advocacy at UCLA and regionally. We work to empower the next generation of human rights lawyers and leaders, generate new thinking on human rights, and engage our students and research to drive positive real world impact.

WITH THANKS TO ADDITIONAL RESEARCHERS

Philip Lockwood-Bean, Brady Mabe, Rie Ohta, Jake Tompkins, Mara Virabov

EXECUTIVE SUMMARY

This report analyzes the relationship between international human rights law and content moderation by social media companies. While states are the primary duty bearers under international human rights law, social media companies have a responsibility to respect human rights, which is heightened during armed conflict. The report looks at how social media companies have dealt with hate speech — which is prohibited under international human rights law — as well as disinformation falling below that threshold, and how their policies measure up to international standards, including the strong protection of freedom of expression under international human rights law. It draws on examples of content posted to four leading social media platforms — VK, Twitter, Facebook and Instagram — prior to, during, and after the period of armed conflict between Armenia and Azerbaijan over the disputed territory of Nagorno-Karabakh in the Fall of 2020. The examples are typical of the misinformation and hate speech seen as part of the information war accompanying the physical conflict. None of this content was subject to moderation by any of the platforms.

Due to the vast amount of content posted during the conflict, we were not able to draw reliable conclusions about the prevalence or character of hate speech or disinformation during this period. We do note, however, that, while we found some instances of hate speech and misinformation posted by Armenian users, it was outweighed by the overwhelming number of posts of that type we encountered from Azerbaijani users.

The report looks at existing platform law to determine how it aligns with international human rights law and whether the platforms should or could have acted to restrict the content. On VK, we discovered an account specifically set up to spread content to help fight the information

war. The account was spreading anti-Armenian hate speech, which is prohibited under international human rights law, and should be criminalized under national law. The relevant VK platform law meets the requirements for restricting expression laid out in Article 19(3) of the International Covenant on Civil and Political Rights and VK should have moderated this content. On Twitter we unearthed an example of disinformation in the form of doctored subtitles which misrepresented the speech of an Armenian official on a linked video. Under the relevant Twitter rules this content should have been subject at least to labeling, but no action was taken. On Facebook we found a video purporting to show Azerbaijani soldiers cutting an ear from an Armenian soldier. We were able to prove that this video was not authentic by comparing it with other similar but authentic content. Under the relevant Facebook rules, this content would not be moderated because the post did not meet the specific purpose of "glorif[ying] violence or celebrat[ing] the suffering or humiliation of others." On Instagram we found a video that had been selectively edited to look as though Armenia was using civilians in armed conflict, posted to an account that falsely claims to be a legitimate Armenian news source, which we disproved. We were able to find the original source of this video to demonstrate that it was not what it seemed. The relevant Instagram rules do not provide for content moderation of this kind of disinformation.

The report observes that content moderation is an extremely difficult task, particularly when the nature of content is not immediately evident. Uncovering instances of manipulation and disinformation is time consuming work. This raises questions about what level of verification and authentication is practical for social media companies to carry out given the vast amount of content posted to platforms every minute of every day. While social media companies have attempted to develop platform law to guide them in their content moderation decision making, that platform law is often unclear or imprecise, and does not always meet the international human rights law threshold to permissibly restrict expression. Content moderation is not something that should be undertaken lightly, given the strong protections for freedom of expression under international human rights law, and the concomitant risks associated with assessing and removing content. Yet even where platform law and international human rights law align, social media companies do not always apply their own policies successfully. In addition, despite strong protections for freedom of expression, there may be policy reasons to moderate content, particularly during an armed conflict where social media posts can influence conflict dynamics on the ground.

The report concludes with some specific recommendations for social media companies to make their internal, self-regulatory policies and content moderation practices more transparent to users (both content creators and content consumers) and better align with international human rights law. These recommendations are intended to be a starting point for deeper discussion on the challenges posed by proliferating harmful content online, particularly during an armed conflict.

INTERNATIONAL COVENANT ON CIVIL AND POLITICAL RIGHTS

ARTICLE 19

1. Everyone shall have the right to hold opinions without interference.

2. Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.

3. The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:

   (a) For respect of the rights or reputations of others;

   (b) For the protection of national security or of public order (ordre public), or of public health or morals.

ARTICLE 20

1. Any propaganda for war shall be prohibited by law.

2. Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.

TABLE OF CONTENTS

INTRODUCTION

In the Fall of 2020, an armed conflict occurred between Armenia and Azerbaijan over the disputed territory of Nagorno-Karabakh/Artsakh ("the conflict"). The physical conflict was accompanied by an online information war where platform users on both sides attempted to influence perceptions of the conflict and the opposing side.[1] This is typical of modern-day armed conflicts and the expanding role of social media, as platforms are utilized to spread fear, hatred, misinformation and disinformation that can directly or indirectly contribute to dire consequences on the ground.[2]

In this report we analyze four examples that typify the kinds of disinformation and hate speech that appeared on social media during the conflict, and were not subject to any moderation by the host platforms. We argue that social media companies must take responsibility for these kinds of speech, and that content moderation decisions should be guided by international human rights law. While states are the primary duty bearers under international human rights law, it is now generally recognized that social media companies, like other businesses, have a responsibility to respect human rights,[3] and the Working Group on the issue of human rights and transnational corporations and other business enterprises suggests that this responsibility is heightened during conflict.[4] International human rights law proscribes propaganda for war and hate speech, under Article 20 of the International Convention on Civil and Political Rights, both of which should be prohibited by law.

In crafting content moderation policies, however, social media companies must keep in mind the strong protections for freedom of opinion and expression provided under Article 19 of the International Covenant on Civil and Political Rights ("ICCPR"), which applies as equally online as offline,[5] even during armed conflict.[6] There is no right to the truth under international human rights law and, consequently, international human

rights law does not prohibit misinformation or disinformation. In fact, the right to freedom of expression includes the right to impart information and ideas,[7] "irrespective of the truth or falsehood of the content,"[8] including when transmitted via the Internet,[9] thereby providing protections for the expression of misinformation and disinformation.

Restriction of any expression, including hate speech, is only permissible under international human rights law if it meets the tripartite test found in Article 19(3) ICCPR: provided by law; for a legitimate purpose, and both necessary and proportionate.[10] The UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Irene Kahn, asserts that "In the light of the fundamental importance of this right to the enjoyment of all other human rights, the restrictions must be exceptional and narrowly construed."[11]

Given the strong protections afforded to expression under international human rights law, moderating online content is a difficult task. Just because content may "shock, offend or disturb"[12] from the perspective of one set of users does not mean that it can or should be restricted or removed, unless it amounts to hate speech or propaganda for war. The same goes for content that is false or misleading.[13] In response to the increasing challenges of moderating content, social media companies have adopted content moderation guidelines — "platform law"[14] — in attempts to protect free speech, regulate user-generated content, and prescribe platform responses to content falling outside of their rules. The result is a confusing patchwork of platform law across social media companies which is opaque and difficult to apply. Whatever the state of individual social media companies' platform law, there is no doubt that reviewing the amount of content posted online is a formidable task. Many major social media companies have thus turned to automation technologies to supplement efforts to flag and filter objectionable and illegal content on their platforms,[15] which is often problematic.[16]

The first part of this report outlines the pertinent international human rights law protections and prohibitions that apply to online content, particularly in a conflict setting. It then explores the platform law of four of the leading platforms in use in the region during the conflict — VK (VKontakte, a Russian online social media and networking service), Twitter, Instagram and Facebook — and examines how those policies align with international human rights law. The second part of the report consists of four case studies of content posted by Armenian and Azerbaijani users to those four platforms during the conflict, identified through open-source investigation ("OSINT") research, that appear to violate platform law but which were not subject to any content moderation. We evaluate whether the identified content could or should have been moderated by the social media company it was posted to in accordance with the relevant platform law and international human rights law. We find platforms failing to moderate hate speech, which is prohibited under international human rights law and should be criminalized under domestic law. We also find that, where the hate speech threshold is not met, freedom of expression often means that content does not warrant moderation under current platform law and international human rights law, even when it may be false or misleading, or undesirable or distasteful to certain users. There may be policy reasons to moderate this type of content, particularly during an armed conflict where expression online may influence conflict dynamics on the ground. In some instances, where content is particularly egregious or manipulated and where moderation is permitted under the rules, social media companies do not always act as prescribed by their platform law. In addition, when content moderation is permitted under platform law, platform law is often not precise enough to enable a restriction of expression in accordance with international human rights law.

The final part of the report provides some specific recommendations for social media companies to make their internal, self-regulatory policies and content moderation practices more transparent to users (both content creators and content consumers) and better align with international human rights law. These recommendations are intended to be a starting point for deeper discussion on the challenges posed by proliferating harmful content online, particularly during an armed conflict.

THE INTERNATIONAL HUMAN RIGHTS LAW FRAMEWORK AND PLATFORM LAW

Social media companies have become central fora for information, discussion, and debate, both during times of conflict and of peace. Since the UN Human Rights Council's adoption of the UN Guiding Principles on Business and Human Rights ("UNGPs") in 2011,[17] businesses, including social media companies,[18] have increasingly come to be recognized as having a responsibility to respect human rights.[19] According to a July 2020 report from the UN Working Group on the issue of human rights and transnational corporations, in a situation of armed conflict, companies not only owe a responsibility to respect human rights in general, but they also come under a heightened responsibility because their business operations may influence conflict dynamics, irrespective of whether the company takes a side in the conflict.[20] The Working Group specifically "names and shames" the tech sector, highlighting that misinformation and hate speech on Facebook fueled the Rohingya genocide in Myanmar, and social media companies should be on notice of this heightened responsibility.[21]

A. REGULATING EXPRESSION UNDER INTERNATIONAL HUMAN RIGHTS LAW

International human rights law provides a framework of standards that social media companies should seek to uphold on their platforms and in content moderation decision making.[22] Article 19 of the International Convention on Civil and Political Rights (ICCPR) protects freedom of opinion and expression including the right to impart information and ideas,[23] even if such information is incorrect.[24] This includes information transmitted via the Internet.[25] Restrictions on expression are permissible under a rigorous, cumulative three-part test enshrined in Article 19(3) ICCPR.[26] The test first requires that the restriction be "provided by law." The second element requires that the restriction pursue either (a) the legitimate ground of respecting the rights or reputations of others or (b) of protecting national order, public order, or public health or morals. The third element requires that the restriction be necessary and proportionate.

Article 20 ICCPR prohibits propaganda for war and hate speech. Its paragraph (1) states that "[a]ny propaganda for war shall be prohibited by law."[27] The travaux préparatoires of Article 20(1) articulate two distinct elements to "propaganda for war."[28] The first element concerns "incitement to war," which the UN General Assembly has interpreted narrowly as a call for conflict or an act of aggression.[29] The second element concerns "the repeated and insistent expression of an opinion for the purpose of creating a climate of hatred and lack of understanding between the peoples of two or more countries, in order to bring them eventually to armed conflict."[30] The second paragraph of Article 20 is commonly referred to as the "hate speech" provision,[31] and provides that "[a]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."[32] The UN Strategy and Plan of Action on Hate Speech has defined hate speech as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor."[33] The Rabat Plan of Action, adopted by the UN Office of the High Commissioner for Human Rights in 2012, sets out a six-part multi-factor test for assessing hate speech and, in turn, the necessity of adopting measures against such speech including: (1) context of the speech, (2) the speaker, (3) whether there is intent to incite, (4) the content and form of the speech, (5) the extent of the speech act, and (6) the likelihood, including imminence, of harm that would result from the speech.[34] The Human Rights Committee has highlighted that even when expression falls into the categories of propaganda for war or hate speech, it may only be restricted in accordance with the cumulative, tripartite test in Article 19(3).[35]

International human rights law does not prohibit misinformation and disinformation. Although a lot of the discourse around social media company moderation efforts is focused on this type of online content,[36] no general consensus has been reached on the definition of these terms, which makes it difficult to address.[37] Whatever the precise content of the terms, Article 19(2) protects freedom of expression even if the expression is false,[38] and so misinformation and disinformation are broadly protected under international

> *"Freedom of expression is not part of the problem, it is the primary means for fighting disinformation"*

human rights law. Irene Khan recognizes that disinformation is a challenge to freedom of expression, but she asserts that "attempts to combat disinformation by undermining human rights are shortsighted and counterproductive. Freedom of expression is not part of the problem, it is the primary means for fighting disinformation" as it allows for alternative viewpoints to be presented and falsehoods and conspiracy theories to be challenged.[39]

B. MODERATING CONTENT UNDER PLATFORM LAW

All social media platforms regulate user-generated content under their own content moderation policies, many of which claim to protect freedom of expression while maintaining a safe space for users.[40] Most of these policies can be found in the companies' Terms of Service or designated rules, but some are scattered elsewhere. This is especially true for Facebook and Instagram policies, an issue that has been raised by the Facebook Oversight Board and which should promptly be addressed.[41]

In this section we provide a brief overview of relevant policies of some of the most popular social media platforms that were in use in the region — VK, Twitter, Facebook, and Instagram — both during the period of the conflict (September to November 2020) where available, and at the time of research in August 2021.[42] What we see is that, while all platforms have developed rules governing the type of content permitted on their sites, some platform law is more transparent and detailed than others. Each platform has a range of moderation mechanisms at its disposal, from labeling to takedowns, but it is not always clear in which circumstance an enforcement mechanism will be applied.

1. VK (VKontakte)

VK strives to achieve a balance between freedom of expression and user safety in accordance with the requirements of Russian legislation.[43] VK states that "transparency and convenient information distribution are at the core" of the product.[44] Essentially, VK aims to strike proportionate responses in its content moderation efforts through its "Safety Guidelines", which chiefly address threats of violence and hate speech.[45] VK enumerates the types of content that users should "refrain" from posting and encourages users to use the "Report button." VK claims that they "address every report, often within several minutes and usually within an hour, [and] If [they] block a profile or community based on a report, [they] notify the user who filed it."[46] In addition to human review, VK "constantly monitors the platform for any harmful content being uploaded…[via] automatic search tools, digital fingerprint technology and neural networks."[47] Rule 7.2.2 of the VK Terms of Service provides that VK may "delete or remove (without giving advanced notice) any Content or Users at is own discretion,… which…infringes these Terms, Russian legislation and/or may infringe the rights of, cause damage to, or threaten the security of other Users or third parties."[48]

2. Twitter

Twitter states that it develops policy "considering global perspectives around the changing nature of online speech, including how [their] rules are applied and interpreted in different cultural and social contexts."[49] With its emphasis on context in its development and enforcement philosophy,[50] Twitter aims to abide by what it refers to as the "Twitter Rules."[51] Twitter's first set of rules are about protecting users' safety.[52] For instance, the "Violent threats,"[53] "Glorification of violence,"[54] and "Abusive behavior"[55] policies aim to promote a healthy dialogue among users by proscribing, *inter alia*, the incitement of violence and calls for serious harm against a group of people. Twitter also has a "Hateful conduct" policy that claims to prohibit the promotion of violence, threats or harrassment on the basis of "race, ethnicity, national origin, caste,

sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease."[56] Twitter's "Sensitive media" policy claims to prohibit "media that is excessively gory or share violent or adult content within live video or in profile of header images."[57]

Twitter also has a set of rules on authenticity.[58] Its policies on "Platform manipulation and spam,"[59] "Impersonation,"[60] and "Synthetic and manipulated media"[61] are all aimed at prohibiting users from deceiving others, whether that be through coordinated inauthentic behavior, the impersonation of someone else, or the dissemination of significantly-altered media.

Twitter's "enforcement philosophy" is grounded in freedom of expression and "promotes counterspeech."[62] Twitter emphasizes that "context matters" in its enforcement decisions, and provides a non-exhaustive list of factors the platform may consider including whether: "the behavior is directed at an individual, group, or protected category of people; the report has been filed by the target of the abuse or a bystander; the user has a history of violating [Twitter's] policies; the severity of the violation; the content may be a topic of legitimate public interest."[63] If a post is found to violate Twitter's policies, Twitter has a range of enforcement mechanisms in place. Tweet-level enforcement can include "labeling a Tweet that may contain disputed or misleading information… Limiting Tweet visibility… Requiring Tweet removal… Hiding a violating Tweet while awaiting its removal…"[64] Twitter also has account-level enforcement policies that may result in "Requiring media or profile edits… Placing an account on read-only mode… Verifying account ownership… Permanent suspension."[65] At the Tweet-level, exceptions may be made if a Tweet is in the public interest. In those instances, Twitter places the Tweet "behind a notice explaining the exception and giving [users] the option to view the Tweet."

3. Facebook

In its Corporate Human Rights Policy, Facebook commits itself to respecting human rights, including the rights set out in the ICCPR.[66] According to Facebook, it implements this commitment by adopting the "responsibility to respect" framework outlined in the UNGPs, specifically by "(1) applying human rights policies; (2) conducting human rights due diligence and disclosure; (3) providing access to remedy; (4) maintaining oversight, governance, and accountability; and (5) protecting human rights defenders."[67] With respect to approach (1), Facebook claims to prioritize human rights in its Community Standards, which govern what user-generated content is or is not allowed on the platform.[68]

Facebook has six sections to its Community Standards.[69] There are multiple policies articulated in each Community Standards section; each begins with a policy rationale setting out the aims of the policy followed by "specific policy lines that outline content that's not allowed; and content that requires additional information to enforce on, content that is allowed with a warning screen or content that is allowed but can only be viewed by adults aged 18 and older."[70] Each policy is quite detailed. Here we focus on those that appear most relevant for content posted in a conflict setting.

Under the "Violence and Criminal Behavior" section, Facebook has policies on "Violence and Incitement,"[71] "Dangerous Individuals and Organizations,"[72] and "Coordinating Harm and Publicizing Crime"[73] which pursue the goal of preventing offline harm by, *inter alia*, prohibiting content that "incites" or "calls for" violence, similar in scope to the Twitter rules. The "Dangerous Individuals and Organizations" policy is particularly concerned with content that supports hate organizations, hateful ideologies, and hate banned entities, delineated into three tiers with different types of enforcement.[74] Tier 1 includes organizations that engage in serious offline harms, for which Facebook removes "praise, substantive support, and representation… as well as their leaders, founders, or prominent members." Tier 2 is geared towards "Violent Non-State Actors," defined as entities that engage in "violence against state or military actors but [who] do not generally target civilians." For this group, Facebook claims to "remove all substantive support and representation of these entities, their leaders, and their prominent members," as well as "praise of the group's violent activities." Tier 3 entities are those that may repeatedly violate Facebook's Hate Speech or Dangerous Organizations policies. This type of content is not

automatically removed to allow room for users to "report on, condemn, or neutrally discuss them or their activities," but users must "clearly indicate their intent when creating or sharing such content." If this is missing, Facebook defaults to removing content. The "Violence and Incitement" policy contains a "misinformation and imminent harm rule" that prohibits "misinformation… that contribute[s] to the risk of imminent violence or physical harm."[75] This policy requires users "do not post" a variety of content, but it is unclear what the enforcement action is if users fail to abide by the policies. This policy also includes information on content that would require additional information or context to enforce.

The section on "Objectionable Content" contains a policy on "Hate Speech" that addresses content targeting a group of people based on a shared identity factor.[76] It also has a "Violent and Graphic Content"[77] policy under which Facebook commits to "remov[ing] content that glorifies violence or celebrates the suffering or humiliation of others." Facebook also notes that "people value the ability to discuss important issues like human rights abuses or acts of terrorism," and suggest that a warning label is more appropriate for this type of content than a takedown. This section has three tiers of prohibited content and information on content that would require additional information or context to enforce.

Under its "Safety" section, Facebook has a policy on "Bullying and Harassment," which distinguishes between private individuals and public figures. The former garners more protection and Facebook professes to remove content that is meant to "degrade or shame." There are seven tiers of content that is not allowed, some of which require self-reporting prior to removal, and additional content that would require more information to enforce on.

Like Twitter, Facebook additionally has Community Standards in place that aim to maintain user integrity and authenticity.[78] For instance, the "Account Integrity and Authentic Identity,"[79] "False News,"[80] and "Manipulated Media"[81] policies seek to combat deception. As with other policies, Facebook lists some guidance on the type of content that is not allowed and for which Facebook will disable accounts, as well as the content for which Facebook would seek further information before taking action.

4. Instagram

Instagram is owned by Facebook.[82] Facebook notes in its Corporate Human Rights Policy that human rights principles guide Instagram's Community Guidelines, which set out Instagram's content moderation rules.[83] These Community Guidelines are relatively thin in comparison to Facebook's policies, but this can be explained by the fact that Instagram's Community Guidelines incorporate many of Facebook's Community Standards by providing external links to them.[84] It appears that Facebook's policies on "Violence and Incitement," "Dangerous Individuals and Organizations," "Hate Speech," "Violent and Graphic Content," and "Account Integrity and Authentic Identity" cover Instagram as well. In addition, Instagram has its own policy on "Reducing the Spread of False Information,"[85] aimed at regulating the spread of misinformation and disinformation.

C. MEASURING PLATFORM LAW AGAINST INTERNATIONAL HUMAN RIGHTS LAW

If we map the international human rights law framework onto platform law, we see that all four social media companies claim to respect freedom of expression. All have detailed rules on hate speech which seem to directly address the type of expression expressly prohibited under Article 20(2). Beyond this, social media companies do not generally follow the same nomenclature of international human rights law in their policies and so there are fewer provisions that specifically address Article 20(1) prohibited content (propaganda for war). Across platforms, we identified no provisions that expressly regulate "incitement to war" or "the repeated and insistent expression of an opinion for the purpose of creating a climate of hatred and lack of understanding between the peoples of two or more countries, in order to bring them eventually to armed conflict."[86]

In a departure from international human rights law which provides explicit protection for expression "irrespective of the truth or falsehood of the content,"[87] the platforms all have policies on misinformation and disinformation which permits moderation of this type of content. How they approach moderation of this type of content varies: for example, Twitter has policies in

place that solely target misinformation in the context of COVID-19, electoral processes, or manipulated media, whereas Facebook has a policy targeting misinformation and false news generally, Instagram has a policy on false news generally, and VK has a vague policy refraining users from posting disinformation.

Whether any moderation (i.e. restrictions) of content under platform law is in line with international human rights law depends on the application of the policy to individual pieces of content and whether any restriction on expression meets the rigorous three-part test enshrined in Article 19(3) ICCPR.[88] Part one requires that the restriction be "provided by law," which must be sufficiently precise and publicly accessible.[89] When considering content posted to social media, the "law" in question is relevant platform law[90] and platform law differs considerably in its precision, depending on the policy in question. The second element of the Article 19(3) test requires that the restriction pursue either (a) the legitimate ground of respecting the rights or reputations of others or (b) of protecting national order, public order, or public health or morals. However, platform law does not always make clear the grounds for its moderation policy. The third element of Article 19(3) requires that the restriction be necessary and proportionate.[91] This is met if the platform law is appropriate to achieve the legitimate ground(s) pursued and there are no alternative, less intrusive enforcement measures available.[92] The fact that all four of these social media companies have varied methods of moderating content means that there are a range of measures available for the platforms to draw on if they determine that moderation is necessary under the relevant platform law. To accord with Article 19(3), content moderation should always pursue the least restrictive means and content removal should always be a measure of last resort.

The platform law of these four social media companies does not yet appear to account for the heightened responsibility to protect human rights during conflict. The UN Working Group on the issue of human rights and transnational corporations outlines three steps that should be taken by businesses in the context of armed conflict: (i) identifying the root causes of tensions and potential triggers, including contextual factors, and the real and perceived grievances that are steering the conflict; (ii) mapping the main actors in the conflict and their motives, capacities, and opportunities to inflict violence; and (iii) identifying and anticipating the ways in which the business's own operations, products, or services impact upon existing tensions and relationships between the various groups and/or create new tensions or conflicts.[93] While it is possible that this is accounted for in internal policies, social media companies should ensure that they are weaving this approach into their moderation decisions and making those policies publicly available so that users can understand how platforms are making decisions.

> *"It is difficult to comprehend how social media companies apply their various content moderation policies in the abstract"*

It is difficult to comprehend how social media companies apply their various content moderation policies in the abstract. The next part of this report looks at some specific pieces of content posted to the platforms during or prior to the conflict in Nagorno Karabakh / Artsakh.

THE ONLINE INFORMATION WAR DURING THE ARMENIA-AZERBAIJAN CONFLICT

On September 27, 2020, a 44-day armed conflict erupted between Azerbaijan and Armenia in Nagorno-Karabakh/Artsakh.[94] The Nagorno-Karabakh Republic, also known as the Republic of Artsakh, is an autonomous state within Azerbaijan's borders,[95] although sovereignty over the territory remains disputed and tensions between Armenia and Azerbaijan have been consistently high since the late 1980s.[96] Within the first few days of the 2020 conflict, tens of thousands of people fled the region. Human rights organizations independently verified that civilians, civilian objects and infrastructure,[97] and medical facilities[98] were targeted in violation of international humanitarian law

and international human rights law,[99] and unlawful weapons were utilized by both sides.[100] There are many reports of Azerbaijan mistreating Armenian prisoners of war.[101] On November 9, 2020, Russia brokered a Trilateral Agreement leading to a ceasefire.[102] This Agreement is not a peace treaty and leaves the status of Nagorno-Karabakh/Artsakh unresolved. Prior to, during, and after the physical conflict the two sides engaged in an online information war, with platform users in both countries using social media companies to fan the flames of the physical conflict.

The Azerbaijani government blocked or slowed access to several social media platforms allegedly to "prevent large-scale Armenian provocations."[103] Platform users inside Azerbaijan were quick to use VPNs in order to circumvent these restrictions.[104] While there is no evidence that Armenia blocked Internet access, it did introduce new censorship measures through amended martial law.[105] These measures forbade the publication of criticism of the government and granted power to the police to levy fines, freeze assets, and demand content removals from the media.[106] In October 2020 alone, Facebook removed 589 Azerbaijani accounts and 7,665 pages from Instagram for exhibiting coordinated inauthentic behavior related to the conflict.[107] Facebook has not published any numbers on the number of Armenian accounts restricted during this period. Several common narratives were advanced across the platforms. From Armenian officials and platform users, there was a tendency to downplay the Armenian military's role in atrocities.[108] It was not uncommon for Armenian officials to project the strength and success of the Armenian military in order to mask the reality that they had suffered major territorial losses;[109] according to Freedom House, the State's information apparatus misled the Armenian public as to genuine developments in the war both on social media and offline.[110] From Azerbaijani platform users, there were prevalent false narratives that Armenia is an aggressor,[111] engaging civilians, child soldiers[112] and foreign fighters in the conflict.[113] Freedom House reports that the Azerbaijani government limited the public's access to unfavorable news and, during the conflict, "much of the media landscape… was dominated by positive coverage of the government and, specifically, the president."[114] Both Armenian and Azerbaijani users attempted to advance competing narratives about historical and cultural ties to the region, asserting that the opposing side had no historical heritage and would destroy the other's cultural property if given the chance.[115]

A. METHODOLOGY

We approached our research into the information war by first identifying popular hashtags, key words, and prominent events/dates used by platform users.[116] We also identified accounts of key State officials, journalists, and individuals who were sharing information about the tensions. We used these search terms and accounts to conduct research on each of the social media platforms in popular use — VK, Twitter, Facebook and Instagram. We used digital verification techniques to confirm details about the posts and account users. Through this process, our team identified more than 250 pieces of content across the various platforms that raised questions and appeared to violate platform law and could potentially be subject to some form of content moderation, but were not subject to moderation by the relevant social media company. Each piece of content identified was then reviewed by four team members and from this four-level review we identified 21 pieces of content that team members believed could potentially be violative of platform law.

From these 21 examples, we have drawn one case study from each of the four platforms that appear to violate platform law.[117] These case studies do not reflect the quantity or distribution of posts during the conflict, but are indicative of the type of content that was being posted online during the relevant period. For example, the first case study looks at anti-Armenian hate speech. While we did find some instances of hate speech and misinformation posted by Armenian users, it was outweighed by the overwhelming number of posts of that type from Azerbaijani users. We included these particular case studies because either they garnered a minimum level of user engagement, the claim in the post spread to other platforms or websites, or both. We made the decision to exclude content that was particularly graphic, inflammatory or offensive, but decided nonetheless to include the hate speech example (which is all of these things) to represent this category of content.

In the following section, we detail our digital verification work on each case and then analyze each piece of content under relevant platform law and international human rights law standards to determine whether they were or

should have been subject to some form of content moderation. We structure our analysis of the content following the Facebook Oversight Board's two-pronged approach of assessing compliance with platform law followed by compliance with international human rights law.[118] We note that content moderation is no easy feat, as the nature of a particular post is not always evident without more in-depth research into both the content and the user. This may not be feasible for social media companies given the amount of content posted to platforms on a minute-by-minute basis. These case studies also show the complexities of moderating content in an environment where posts are being reposted and shared on different platforms.

CASE STUDY 1: POLYGON AZERBAIJAN — APPLYING VK PLATFORM LAW

This case study examines posts from a popular VK account — Polygon Azerbaijan — that creates and spreads content to help fight the "information war."[119] We found a large volume of extreme content targeting Armenians from Azerbaijani accounts on VK but did not find examples of hate speech directed at Azerbaijanis and posted to Armenian accounts on this platform. An investigation into hate speech on Twitter by other researchers found similar results.[120]

Figure 1.1 The Polygon Azerbaijan YouTube banner with the VK logo inset on the right.

Polygon Azerbaijan is a military and defense-themed social media brand that has produced unique satirical and developing-events content throughout the conflict. The brand was created and is operated by the pseudonymous independent journalist Hans Kloss,[121] and its primary VK account[122] has 10,542 Followers. Polygon Azerbaijan also maintains a YouTube channel[123] (793 Subscribers) but does not have a Twitter, Facebook, or Instagram presence.[124] Despite attempts to identify him, we do not know Kloss' true name. He and his collaborators are consistent with their use of the pseudonym in interviews and media productions. He is careful to obscure his face when he appears in front of the camera. Though Kloss often indicates that the terms "Hans Kloss" and "Polygon Azerbaijan" are copyright protected, WIPO[125] searches produce no records of these brands. Searches of other social media platforms, public records, media reports, and image databases produced no definitive information on Kloss' true identity. Kloss frequently collaborates with former Azerbaijani government official, military analyst, and Azerbaijani state television host Heydar Mirza[126] on the video journalism projects Caliber.az[127] and RADIUS.[128] These productions are often linked or reposted on the Polygon Azerbaijan sites.

Polygon Azerbaijan content often referred to Armenians as vermin,[129] and graphics featuring rats as Armenian soldiers were frequently posted on Polygon Azerbaijan in response to announcements of the death or serious injury of Armenian military personnel. Here we provide just a few examples

FIGURE 1.2 *Translation: Category • here I come! (updated) | Two servicemen of the occupation contingent of the Armenian Armed Forces on the Territory of Azerbaijan were blown up by a mine near their positions in the 7th so-called "defensive area" (military unit: #38862) in Tonashen. | Khoren Anushavanovich Shagulyan, conscript soldier (lesion and further amputation of the lower limb) | Levon Hayriyan Khachikovich, conscript soldier (wounded in the face and neck) | P.S. Someone, wake up Shushan Shagenovna already, otherwise her infernal snoring can already be heard in Barda... | Hans Kloss © | Based on: mobile application here I come!*

FIGURE 1.3 The video titled "Deathrat: stationary dance of the Armenian invaders..." appears to be low-flying drone footage of a purported Armenian trench line and dead Armenian soldiers (circled in red by the content creator).

Translation: category • here I come! | Deathrat: stationary dance of the Armenian invaders... | Hans Kloss © | Based on: mobile application here I come!

of that content, dating from July 2020 to March 2021. Most posts and comments that appear on Polygon Azerbaijan's accounts are written in the Russian and Azerbaijani languages. We used Google[130] and Yandex[131] AI translators to translate Russian into English. We then verified those translations with a native Russian speaker and adjusted for accuracy and comprehensibility. Azerbaijani translations were created using Google and Yandex alone. All of the content analyzed below was originally written in Russian. An English translation is provided for post content as well as any wording that appears on a posted graphic.

The post at Figure 1.2 was posted on July 2, 2020[132] — almost two weeks before the July 2020 clashes. Kloss constructs the graphics using photos of injured or dead rats available on the Internet and adds images of clothing, weapons, and other details associated with Armenian military members. In the bottom right of the image, there appears the national emblem of Armenia, upside down and with a bullet hole, atop the Armenian national flag. This post received 2,500 views, 91 likes, 5 comments, and 9 shares as of September 2, 2021.

The most common appellation, "rat," was often applied to dead Armenian soldiers, as is the case in Figure 1.3. This post was made on October 11, 2020[133] — during the height of the Fall conflict. It received 23,000 views, 506 likes, 61 comments, and 21 shares as of September 2, 2021.

Figure 1.4 provides another example of Polygon Azerbaijan referring to dead Armenian soldiers as rats (Russian: крысы). This content was posted on October 18, 2020[134] — during the height of the Fall conflict. The post received 17,000 views, 531 likes, 58 comments, and 16 shares as of September 2, 2021. In this case, the soldiers are claimed to have been killed near Fuzuli.[135] In addition, the post also describes Armenians as носатие, which means nosy or big-nosed. In some cases, Armenians express pride or poke gentle fun at their noses;[136] however, Polygon Azerbaijan draws attention to this feature in a mocking and demeaning manner that is reminiscent of prominent anti-Semitic stereotypes.[137]

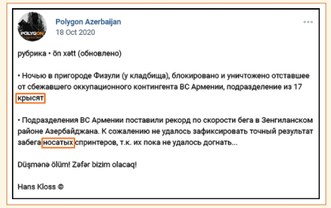

FIGURE 1.4 This content describes the purported combat deaths of Armenian soldiers during Azerbaijan's military takeover of the city of Fuzuli. Objectionable terms are highlighted by colored boxes in the post and are underlined in the English translation.

*Translation: Category * front line (updated) | At night, in the suburb of Fuzuli (near the cemetery), a unit of 17 baby rats that had fallen behind the escaped occupation contingent of the Armenian Armed Forces was blocked and destroyed | Units of the Armenian Armed Forces set a record for running speed in the Zengilan region of Azerbaijan. Unfortunately, it was not possible to record the exact result of the race of big-nosed sprinters, because they have not yet been able to catch up... | Death to the enemy! Victory will be ours! | Hans Kloss ©*

FIGURE 1.5 In addition to the Armenian military equipment details in image, the author also includes a skull and bullet-pocked Armenian national emblem. Below the image is a linked music file with text translating to "Last Dance of the Rat — Black Trumpet".

FIGURE 1.6 The author has added a weapon and Armenian military cap to the image as well as a skull and the English word "Destroyed". The name of Aghasi Hovhannisyan appears in a black box at the top of the image. At the bottom is a linked music file with text translating to "Last Dance of the Rat — Black Trumpet."

Figure 1.5 is another example of the satirical dead rat post. The post was made on December 29, 2020 after the end of the Fall conflict.[138] This post received 6,700 views, 254 likes, 19 comments, and 14 shares as of September 2, 2021. The uniform, equipment, and symbols of the Armenian state evoke from the viewer a sense of dehumanization and celebration of the deaths of Armenian service members. Polygon Azerbaijan's mocking post is a response to a report[139] that two Armenian service members were killed when their vehicle ran off the Gorus-Kapan road.

The post at Figure 1.6 was posted on March 23, 2021,[140] more than four months after the end of the Fall conflict. It received 5,200 views, 131 likes, 18 comments, and 9 shares as of September 2, 2021. The soldier referenced in Figure 2.10, Aghasi Hovhannisyan, became lost with another soldier during a severe snowstorm and died on March 21st or 22nd.[141] The exclamation in the graphic refers to Zangazeur, which is the name of both the mountain range in which the soldier died and an important battle in Armenian history.[142] Its inclusion in the graphic may have a double-meaning intended to mock Armenian military pride.

These Polygon Azerbaijan VK posts and images have been reposted to Twitter accounts,[143] message boards,[144] and independent news sites.[145] Figures 1.5-1.6 show that Polygon Azerbaijan continued to post content derogatory of Armenians and the Armenian military well after the end of the Fall 2020 conflict. This reflects the continued enmity directed at Armenians online and the danger of renewed hostilities between Azerbaijan and Armenia.[146]

As mentioned, Polygon Azerbaijan is connected to a larger visual media network, which includes the video journalism projects Caliber.az as well as RADIUS, a program which ran on Azerbaijani State television.[147] Kloss partnered with Azerbaijani state television host Heydar Mirza in developing both programs.[148] Heydar Mirza is the professional face of this media network with a background in military and political analysis. He has a Ph.D. in International Relations from Freie Universität Berlin and worked as a strategic studies analyst under the Azerbaijani president for eight years.[149] In the partnership between Kloss and Mirza, there is a kind of official legitimacy lent to the type of content posted on Polygon Azerbaijan.

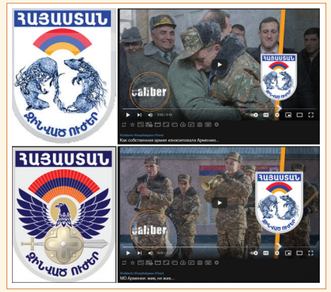

FIGURE 1.7 The top left panel is an enhanced version of the emblem that appears on the preview screen of the two Caliber.az YouTube videos[150] shown in the two panels on the right. The bottom left panel is the official emblem of the Armed Forces of Armenia. Both emblems on the left contain the same Armenian wording which translates to "Armenian Armed Forces." The video on the top right was posted March 17, 2021 and has received 12,376 views, 103 comments, and 1,100 upvotes to 18 downvotes. The video on the bottom right was posted March 23, 2021 and has received 11,114 views, 71 comments, and 917 upvotes to 12 downvotes.

Polygon Azerbaijan sites promote both Caliber.az and RADIUS content, which tends to be free of the explicitly derogatory elements that pervade Kloss' own site. Still, even these more "professional" sites occasionally include references to the extreme narratives promoted on Polygon Azerbaijan as demonstrated in Figure 1.7.

1. Compliance with Relevant Platform Standards

The VK Terms of Service[151] state that a user is prohibited from making available any content that: "contains threats or calls to violence, including ones made implicitly; praises or encourages violent actions or discredits; insults; defiles one's honor, dignity or business reputation";[152] "contains scenes of inhumane treatment of animals";[153] "propagates and/or incites racial, religious, or ethnic hatred or hostility, including hatred or hostility towards a specific gender, orientation, or any other individual attributes or characteristics of a person[…]";[154] "propagandizes and/or contributes to racial, religious, ethnic hatred or hostility, propagandizes fascism or racial superiority";[155] or "is of fraudulent nature".[156]

Under its Safety Guidelines,[157] VK explicitly prohibits users from spreading hate speech or to otherwise "victimize or belittle an individual or group of people based on religion, culture, race, ethnicity, nationality, sexual or gender identity, developmental differences, illness, etc." The platform claims to block accounts that spread content that contains "verbal assertion[s] of superiority of some groups over others to rationalize violence, discrimination, segregation, or isolation on the basis of religion, ethnicity, nationality, sexual or gender identity, developmental differences or illness" even in cases in which the content is posted as a joke or meme.[158] Qualifying assertions include "comparing a specific group of people to insects, filth, subhumans, inferior types, and other such language."[159] VK further asks users to refrain from posting content that glorifies violence, depicts physical harm, or contains disinformation.[160] The platform does acknowledge the importance of context in assessing prohibited content, and moderators are instructed to look for evidence that content was posted maliciously.[161] Content that violates these policies and is maliciously posted may be deleted or result in a user losing their account; repeated violations may result in a permanent ban from the platform.[162]

Polygon Azerbaijan's content is consistent with multiple categories of VK's prohibited content under VK's Terms of Service. Under the platform's policies, each of the example posts should have been removed, and there is a strong case that the entire account should be taken down.

The Polygon Azerbaijan posts presented here propagate and propagandize racial and ethnic hostility or hatred in violation of VK's Terms of Service (¶¶ 6.3.4e-f). The posts in Figures 1.3-1.6 all include graphic or written comparisons of Armenians to rats, which falls under VK's description of hate speech in its Platform Standards.[163] In each of these posts, the comparison to vermin is made within the context of the death or injury of Armenians—posts in Figures 1.4 and 1.7 depict actual physical violence to people who are described as Armenian service members. The inclusion of this context plausibly may be said to glorify violence or incite hostility as the depiction or description of violence in these posts is celebratory, particularly for the posts in Figures 1.4, 1.5, and 1.7.

The posts in Figures 1.5, 1.6, and 1.7 also describe Armenians by the physical size of their noses. The negative stereotype equating this physical characteristic with the Armenian nationality, ethnicity, and race and its use in content that promotes or celebrates physical violence targeted at Armenians, violate VK's Platform Standards by asserting the superiority of Azerbaijanis over Armenians to rationalize violence and discrimination on the basis of nationality and ethnicity.[164]

According to VK policies, the satirical nature of many of the example posts is no defense against moderation. Under VK's maliciousness test, Polygon Azerbaijan's posts demonstrate (1) animosity based on certain characteristics or differences (e.g. "big-nosed"); (2) offensive behavior, contempt toward other people's values or views; and (3) expression of personal superiority, accompanied by a baseless and unfair attitude toward a specific individual or group of people (e.g. Armenian soldiers killed in combat).

Given the extremity of the content available on the Polygon Azerbaijan account, particularly as it promotes hostility and hatred toward Armenians of ethnic and national identities, VK's policies require that, at the least, all eight offending posts be removed, and there is a persuasive case that the owner should lose the account, be permanently banned, or both. Hans Kloss and Polygon Azerbaijan likely qualify for these severe sanctions under VK's policy against posting hate speech as well as under its policy on repeat violators.

2. Compliance with international human rights law

Businesses, including social media companies, should respect human rights. This means that they should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved.[165] Hate speech is a prohibited form of expression under Article 20(2) ICCPR. Any restriction of hate speech must meet with the cumulative requirements of Article 19(3) ICCPR: (i) that any restriction be provided by law, (ii) that the restriction pursue either (a) the legitimate ground of respecting the rights or reputations of others or (b) of protecting national order, public order, or public health or morals, and (iii) that the restriction be necessary and proportionate.[166] The posts presented from Polygon Azerbaijan qualify under the UN definition of hate speech as expressive content that "uses pejorative or discriminatory language with reference to a person or a group on the basis of [...]" their ethnicity, nationality, race, or other identity factor.[167] Further, this content is prohibited by Article 20(2) ICCPR which outlaws any "advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence."[168] In particular, Polygon Azerbaijan's use of the offending terms and images in combination with military content promotes hostility and potentially violence toward Armenians.

VK's policy on hate speech is provided by law as required by Article 19(3) as it is contained in the Platform Standards section of their publicly available Safety Guidelines document.[169] The examples contained in the Safety Guidelines provide users with a clear understanding of particular categories of prohibited content. VK policies make clear that the content in these posts is prohibited and should be moderated. However, the outcome for other content — that which is not clearly encompassed by the examples provided in the Platform Standards — may be more difficult to predict. Therefore, we recommend that VK update its Platform Standards policy to make explicit its definition of hate speech and clarify the criteria it uses to identify and moderate all forms of hate speech.

VK's Platform Standards document describes its hate speech policy as intended "to ensure a safe environment for all."[170] Although this is vague, it likely comports with the Article 19(3)(a) restriction in favor of respecting the rights or reputations of others.[171]

When considering moderation of hate speech in particular, the six-part test of the Rabat Plan of Action[172] provides a robust deliberative, high-threshold framework to assess whether moderation actions are necessary and proportionate to the content in question, considering the totality of the circumstances. This was originally developed to assess the necessity of adopting criminal measures against hate speech,[173] but has since been extended beyond expression that is criminalized.[174] This test includes (1) social and political context, (2) the speaker's position or status, (3) the intent of the speaker, (4) the content and form of the statement, (5) the extent of its dissemination, (6) the likelihood of harm, including imminence.

(1) **Context:** A state of open, large scale armed conflict existed between Armenia and Azerbaijan between the months of September and November 2020. For months before and in the many months following, almost continuous armed confrontations occurred between the countries resulting in combat casualties. The cease-fire negotiated in November 2020 remains uneasy, and current conditions on the ground resemble to some degree the build up to war that existed in July and August of 2020. Meanwhile, the Azerbaijani President, Ilham Aliyev, has celebrated his country's victory in a manner that is meant to humiliate Armenia[175] and has used dehumanizing language to describe Armenians.[176] It is in this context of intense international armed conflict and national and ethnic enmity that Polygon Azerbaijan's posts appeared, presenting Armenians as dangerous and subhuman at a time and in a region where there remains a high probability

of violence. These posts also reflect a broader strategic use of hate speech by Azerbaijan's political leaders. Consequently, the social and political context increases the severity of the content and favors moderation.

(2) **Status of the speaker:** the speaker Hans Kloss, the account owner, purports to be a military expert who has worked for and appeared on Azerbaijan state television. He also collaborates with Heydar Mirza, a television personality and former national security official in the Azerbaijani President's administration. Kloss' content is popular with his intended Russian- and Azerbaijani-speaking audience, and his posts are often reposted to other platforms. As a popular and politically-connected figure within Azerbaijani media, Kloss' status favors moderation.

(3) **Intent:** a speaker's intent to incite is signaled by the deliberate coupling of hate speech and depictions of violence against a group. The posts from Polygon Azerbaijan combine hate speech with visual depictions or written descriptions of physical violence committed against members of the Armenian military. In some cases, this includes graphic photos or videos showing dead Armenian service members. Further, Mirza and Kloss produce a significant amount of content that would appeal to military members or military hobbyists,[177] and Polygon Azerbaijan posts are reposted to military-themed message boards.[178] Polygon Azerbaijan posts celebrate the killings, woundings, and accidental deaths of Armenian service personnel with dead rat memes and dehumanizing language. These elements suggest intent to incite and favor moderation.

(4) **Content and form:** The Polygon Azerbaijan posts are one-sided, were created by a single person, and most present no intellectual argument. The account is hostile to alternative views; Armenian perspectives are only featured in Polygon Azerbaijan posts as an opportunity for mockery or derision, not legitimate points of argument. The posts appear targeted to provoke strong emotional reactions from the audience as well as the targets. The substance and form of the speech favor moderation.

(5) **Dissemination:** Content is more liable to be moderated when it is part of a continuing campaign and appears on a publicly-accessible website where it is encountered by thousands of people who have access to the target population.[179] Polygon Azerbaijan is on the VK social media platform and its content is available to anyone with a VK account.[180] The posts that presented here are not isolated examples but contain themes and language that are repeated many times in content that spans from well before the Fall conflict to at least summer 2021. Many of the examples here have been reposted to news sites,[181] other VK accounts,[182] Twitter accounts,[183] and standalone sites.[184] The collaboration between Kloss and Mirza extends the reach of this content even further and into more traditional and official Azerbaijani media, beyond VK. The Caliber.az project is ostensibly a video journalism project that reports on the ongoing conflict;[185] however, even on Caliber.az videos the Armenians-as-rats messaging appears. Because the content is widely accessible online and much of the audience has access to the target population, the extent of the speech increases its severity and favors moderation.

(6) **Likelihood of harm:** The Polygon Azerbaijan posts promote narratives that suggest Armenians are (1) dangerous and deserve to be the targets of violence; (2) subhuman and do not deserve dignity or safety (e.g., Armenians are "rats", "creatures"); or (3) incompetent and draw injury upon themselves (e.g., mocking soldiers who become lost in a snowstorm). These narratives — which mirror those promoted by media outlets prior to and during the Rwandan genocide[186] — in combination with the similarly hateful and violent rhetoric promoted by Azerbaijan's political establishment, increase the risk of violence, thereby increases the severity of the speech and favors moderation.

An analysis of each of the six Rabat factors shows that this content on the Polygon Azerbaijan VK account meets the threshold for criminalisation under national law, and should be removed from the site. Given the extreme nature of these posts there is a case to be made that the entire account should have been blocked: VK states that it blocks accounts in cases of the most egregious speech or of repeat violators of content policies, otherwise moderation may

be limited to warning or post deletion. At minimum these pieces of content on the Polygon Azerbaijan VK account should have be blocked or deleted, and in failing to do so VK failed to apply its own platform law and to respect human rights.

CASE STUDY 2. AZERBAIJAN MFA TWEET: APPLYING TWITTER PLATFORM LAW

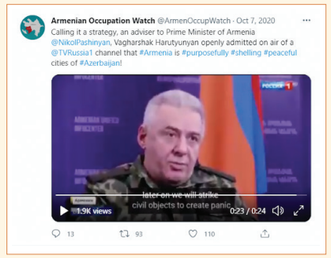

FIGURE 2.1 Screenshot of the video posted to Twitter by the "Armenian Occupation Watch," which received 1.9k views as of August 13, 2021.

During the conflict, it was not uncommon for Azerbaijani public officials to make statements on Twitter claiming that attacking civilian populations was a standard practice of the Armenian forces.[187] We did not find examples of similar content posted by Armenian officials. On October 7, 2020, a video was posted on Twitter by the user Armenian Occupation Watch (@ ArmenOccupWatch). According to the bio, this account is managed by the Ministry of Foreign Affairs of Azerbaijan ("MFA"). The video showed a clip of Vagharshak Harutyunyan, a former advisor to the Armenian Prime Minister during the conflict, speaking Armenian with English subtitles (see Figure 1.1).[188] The post claimed that Harutyunyan was explaining that Armenia's is "#purposefully #shelling #peaceful cities of #Azerbaijan." This video had received 1.9k views as of August 13, 2021. The Twitter post received 110 likes, 93 retweets, and 13 comments as of the same date.

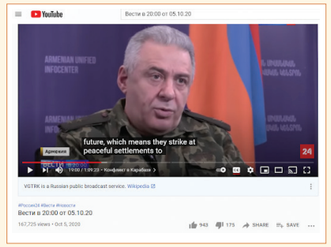

FIGURE 2.2 The original clip on YouTube, posted on October 5, 2020. Harutyunyan is talking about "they," referring to the Azerbaijanis (translated using YouTube's auto-generated English translation and confirmed by members of our team).

The same clip was posted by another Twitter user (@Shahlam_) on October 9, 2020, although the clip did not include the English subtitles like the video posted by @ArmenOccupWatch.[189] The @Shahlam_ Tweet revealed that the clip was from YouTube and featured the title of the video — "Вести в 20:00 от 05.10.20." We were able to use this title to identify the original video on YouTube from which the clip of Harutyunyan was taken. That video was posted to YouTube on October 5, 2020 by Россия-24, a Russian news channel (see Figure 1.2). By listening to the original clip with YouTube's automated English translation turned on, we were able to identify that the English subtitles in the @ArmenOccupWatch post are misleading: Harutyunyan is not speaking about Armenia's military tactics, he is speaking about the military tactics of Azerbaijan. This was further confirmed by members of our team who are familiar with the language being spoken in the video.

1. Compliance with Relevant Platform Standards

Twitter's "Synthetic and manipulated media" policy provides that users "may not deceptively promote synthetic or manipulated media that are likely to cause harm."[190] Additionally, Twitter "may label Tweets containing synthetic and manipulated media to help people understand their authenticity and to provide additional context."[191]

Twitter enforces its "Synthetic and manipulated media" policy in accordance with a three-part test that asks (i) whether the content was significantly and deceptively altered or fabricated, (ii) whether the content was shared in a deceptive manner, and (iii) whether the content is likely to impact public safety or cause serious harm.[192] Different enforcement actions are taken depending on how many of these criteria the content in question satisfies.[193] In assessing criterion (i) on whether the media has been deceptively altered, Twitter expressly notes in its policy that it considers whether modified subtitles have been added.[194] Twitter provides it is most likely to take strong moderation action against media that has been significantly altered (e.g., spliced and reordered or slowed down to change its meaning), however under Twitter's rules content containing subtler forms of manipulation, such as presentation with false context, may be labeled or removed on a case-by-case basis.[195]

When assessing criterion (ii) on whether the content was shared in a deceptive manner, Twitter considers whether the content suggests a deliberate intent to deceive people, taking into account the text of the Tweet and information on the profile of the account sharing the media.[196] Finally, in analyzing criterion (iii) on whether the content is likely to impact public safety or cause serious harm, Twitter considers, *inter alia*, threats to the physical safety of a person or group and the risk of mass violence or widespread civil unrest.[197]

Here, the MFA Tweet added inaccurate subtitles to the video featuring the Armenian official, which satisfies criterion (i). As the inaccurate subtitles had to be created, it seems likely that this deception was intentionally shared, which satisfies criterion (ii). Whether or not criterion (iii) is satisfied is more complex. Broadly, it could be argued that because the MFA's post claims that Armenia deliberately strikes at civilians, it was created in order to incite panic and to stir up further animosity between the two sides, which could impact public safety or cause serious harm if content consumers used this information as a pretext

to respond to the perceived Armenian action. Twitter considers the time frame within which the content may be likely to impact public safety or cause serious harm,[198] so the potential for serious harm is supported by the fact that the post was made on October 7, 2020, just 10 days after the physical conflict materialized. Twitter identifies some specific harms included in criterion (iii) including "threats to the physical safety of a person or group; risk of mass violence or widespread civil unrest; threats to the privacy or ability of a person or group to freely express themselves or participate in civic events." While this list is not exhaustive, it does not appear that this Tweet would cause serious harm of this nature. Consequently, we conclude that although the MFA's post does heighten tensions between Armenia and Azerbaijan, the link between this content and serious harm is perhaps too tenuous (criterion (iii)).

However, under Twitter's policy a post does not need to satisfy all three criteria. For content that satisfies criteria (i) and (ii),[199] Twitter's policy states that it is "likely to be labeled." No such action was taken against this post.

In its policy, Twitter states that it may use its own technology or receive reports through partnerships with third parties in order to determine if media has been deceptively altered.[200] It may be the case that Twitter's algorithms and third-party partners did not detect the MFA's post, so we recommend that Twitter's algorithms and partners prioritize synthetic and manipulated media shared in the context of an armed conflict.

2. Compliance with international human rights law

In order for content moderation to respect human rights law, it needs to meet the cumulative criteria of Article 19(3) ICCPR: (i) that any restriction be provided by law, (ii) that the restriction pursue either (a) the legitimate ground of respecting the rights or reputations of others or (b) of protecting national order, public order, or public health or morals, and (iii) that the restriction be necessary and proportionate.

By setting out a three-part test that delineates Twitter's approach to synthetic and manipulated media, Twitter's policy is formulated with sufficient precision to allow individuals to foresee the types of conduct that are prohibited under the policy and thus accords with the "provided by law" component of Article 19(3). It is especially helpful that Twitter includes a table in its policy, which explains the enforcement action Twitter will take depending on how much of the criteria in its three-part test is satisfied. This policy can be easily located on Twitter's Help Center, which means that the policy has also been made accessible to the public. However, Twitter could improve its transparency by being specific about when labels will be applied, and the content of those labels, rather than just providing that labeling "is likely."

Twitter's policy rationale for its "Synthetic and manipulated media" policy states that: "You may not share synthetic, manipulated, or out-of-context media that may deceive or confuse people and lead to harm."[201] This is vague. It could be read as aimed at protecting public order and, in more extreme circumstances, national security under Article 19(3)(b), but it is not explicit. To be more transparent, Twitter should make the human rights that it seeks to protect more explicit on the face of the policy in order to provide greater clarity.

Whether a restriction on this expression would comport with the necessity and proportionality test of Article 19(3) depends on the type of restriction applied. Since the MFA is a department of the State of Azerbaijan and the content it posted relates to the conflict, the post is political speech and is subject to particularly strong protection under the right to freedom of expression.[202] We find that removal of the MFA Tweet would not be necessary and proportionate because, as discussed above, we do not believe that it reaches the "serious harm" threshold expounded by Twitter that would warrant its removal.[203] The objective could be met via the less restrictive action of labeling the MFA Tweet as manipulated and inaccurate, as permitted under Twitter's policy, which would be necessary and proportionate in this instance. This conclusion is further bolstered by the fact that the real video, which excludes the MFA's modified subtitles, is freely available on YouTube and can be used to counter-message the MFA Tweet.

Consequently, we conclude that this Twitter policy respects human rights as articulated in Article 19(3) of the ICCPR, although it failed to uphold its own platform law in not labeling the MFA Tweet.

Twitter could improve its approach to misleading content by adding links to trustworthy sources as part of its enforcement response.

CASE STUDY 3. ARMENIA-ARTSAKH AWARENESS CENTER (AAAC) POST — APPLYING FACEBOOK PLATFORM LAW

FIGURE 3.1

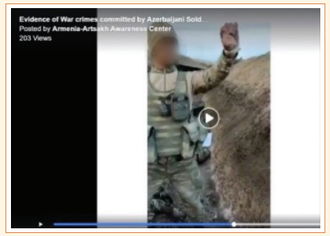

FIGURE 3.2

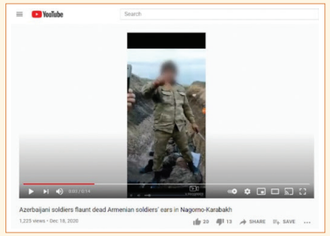

FIGURE 3.3 The video was uploaded to YouTube on December 18, 2020, amassing 1,236 views as of August 12, 2021.

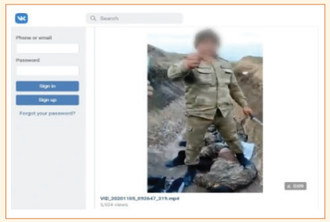

FIGURE 3.4 The video was uploaded to VK on November 5, 2020, amassing 5,924 views as of August 3, 2021.

On November 4, 2020, a Facebook user by the name of the "Armenia-Artsakh Awareness Center" ("AAAC") posted a video allegedly showing Azerbaijani forces cutting off the ears of an Armenian soldier (see Figures 3.1 and 3.2).[204] This account was created on October 6, 2020, in the midst of the conflict, and was actively posting pro-Armenia/anti-Azerbaijan content up until April 1, 2021. It has amassed 1,553 followers as of August 12, 2021. We were able to trace ownership of this account to a private citizen based in Los Angeles. Many of this person's posts on the AAAC's Facebook page, including the ear-cutting video, contain pleas for donations to the ArmeniaFund, a nonprofit organization dedicated to serving the humanitarian needs of Armenia and Nagorno-Karabakh/Artsakh.[205]

By performing a reverse image search on the screenshots at Figure 3.1-3.2 taken from the video using Google Images and Yandex Images, it became clear that this video had spread to other social media platforms including YouTube, VK, and Twitter, as evidenced in Figures 3.3-3.5.

An OSINT investigation carried out by Amnesty International's Crisis Evidence Lab confirmed the authenticity of twenty-two videos depicting gruesome mutilations and extrajudicial executions of enemy soldiers by both Armenian and Azerbaijani forces during the conflict.[206] Ear-cutting by both Armenian and Azerbaijani perpetrators was thus not uncommon during the conflict,[207] but comparing the Facebook video posted by the AAAC with videos and images authenticated by Amnesty International raises questions of the authenticity of the AAAC video. Journalist Jake Hanrahan posted authentic images of Azerbaijani soldiers cutting off the ears of Armenians to Twitter.[208] Of note is the fact that blood is visibly apparent in these images, a feature that is missing in the AAAC Facebook video. On the contrary, the "ear" in the AAAC Facebook video is scrupulously clean, as is the knife that was used to allegedly cut off that ear. No blood is detectable anywhere in the AAAC Facebook video. In addition, the "ear" in the AAAC Facebook video does not

FIGURE 3.5 A screen capture from the video was also shared on Twitter on December 17, 2020, garnering 281 retweets and 258 likes as of August 3, 2021.

observably look like an ear at all. Rather than appear life-like, the "ear" in the AAAC Facebook video is nearly flat and surrounded by flappy excess "skin." Moreover, Amnesty International points out that the uniforms of Azerbaijani soldiers are typically marked by the Azerbaijani flag on the shoulder and a patch with the soldier's blood type on the sleeve.[209] These distinctive features are missing from the AAAC Facebook video. We conclude that this video is not authentic.

1. Compliance with Relevant Platform Standards

At the time of the conflict, Facebook's "Violent and Graphic Content" policy prohibited posting videos of people or dead bodies in non-medical settings if they depicted dismemberment.[210] Facebook's rationale for its "Violent and Graphic Content" policy stipulates that Facebook "remove[s] content that glorifies violence or celebrates the suffering or humiliation of others because it may create an environment that discourages participation [on Facebook]."[211] The policy rationale further provides that Facebook "allow[s] graphic content (with some limitations) to help people raise awareness about these issues."[212] For content that falls within the policy, Facebook indicates that it will "include a warning screen so that people are aware that the content may be disturbing."[213] In this case, the AAAC's post does not glorify violence or celebrate the plight of the Armenian soldiers; instead, it does the opposite by calling out the alleged Azerbaijani military violations of the treatment of prisoners of war. Consequently, the post did not violate Facebook's "Violent and Graphic Content" policy and, therefore, was properly not the subject of moderation under this policy.

Facebook's "Violence and Incitement" policy prohibits posting "[m]isinformation and unverifiable rumors that contribute to the risk of imminent violence or physical harm."[214] According to an update made to this policy on Facebook Newsroom a month before the conflict broke out, misinformation that does not attain the imminent harm threshold but is rated false by third-party fact-checkers will be downranked in the News Feed.[215] Here, context matters:[216] This post was made five days before the conflict came to an end when there was still uncertainty over whether

(temporary) peace could ever be accomplished between the two sides. In addition, the post casts Azerbaijani soldiers as war crime perpetrators, which could feed into negative generalizations about Azerbaijanis. On the other hand, the post did not use derogatory language against Azerbaijanis[217] and did not enjoy substantial user engagement. Although the ear-cutting video was circulated across different platforms, the post made by the AAAC was not liked, shared, or commented on very much, which points to a lack of imminent harm. Thus, under Facebook's policy, the post should not have been removed, but it could have been downranked if rated false by third-party fact-checkers. There is nothing to suggest that the post was indeed rated false by third-party fact-checkers. so we recommend that Facebook provide greater transparency on how it partners with third-party fact-checkers and how its algorithms and human review processes work.

2. Compliance with international human rights law

Whether this content can be restricted under international human rights law depends on the satisfaction of the cumulative three requirements of Article 19(3) ICCPR: (i) that any restriction be provided by law, (ii) that the restriction pursue either (a) the legitimate ground of respecting the rights or reputations of others or (b) of protecting national order, public order, or public health or morals, and (iii) that the restriction be necessary and proportionate.

Facebook could have moderated this content in line with the requirements of Article 19(3). Its "Violent and Graphic Content" policy meets the first requirement of Article 19(3) as it is sufficiently precise and the policy is publically available. It appears as though the policy is aimed at "respect[ing] the rights or reputation of others," and so satisfies the second requirement. The method of moderation provided under the policy is limited to providing a warning, which is not particularly restrictive and so would likely meet the necessity and proportionality test: the third requirement of Article 19(3). However, the post would not be moderated under the terms of the policy, as it is specifically directed at content that "glorifies violence or celebrates the suffering or humiliation of others," which this post did not do. We note that this might be an unsatisfactory outcome, as it allows violent and graphic content which is posted for other purposes, such as this.

Facebook's "Violence and Incitement" policy and "misinformation and imminent harm" rule are publicly accessible on Facebook's Transparency Center, and appears to meet with the legitimate ground required by Article 19(3) as Facebook states that its goal is to prevent offline harm and threats to public safety.[218] However the rule contains two major deficiencies. First, the rule does not define "misinformation," leaving open to interpretation of the types of prohibited content. This is a problem that has been raised by the Facebook Oversight Board more than once.[219] Second, the rule does not explain what the "imminent harm" threshold is. Consequently, this rule does not fulfill the "provided by law" criteria of Article 19(3) and, as the test is cumulative, content moderation under the policy would not be permitted by international human rights law.

Even if the Article 19(3) test were met for the "misinformation and imminent harm" rule, Facebook's policy, which permits downranking the content if it is rated false by third-party fact-checkers, is not sufficient. The post was made during the conflict, which should have triggered Facebook's heightened responsibility to respect human rights under the UNGPs. Therefore, just as the FBOB has recommended that Facebook prioritize referring content to its fact-checkers when the content concerns a public position on debated health policy issues, particularly in the context of a pandemic,[220] we recommend that Facebook prioritize referring content that makes dubious claims about an armed conflict to its fact-checkers. Also, we recommend that Facebook expand its policy to include alternative, less intrusive measures to downranking, such as affixing a label that warns users of misinformation and/or directs users to trusted sources of information. These measures could protect users' freedom of expression, while also allowing other users to explore alternative sources of trustworthy information.

We conclude that Facebook's "Violence and Incitement" policy and "misinformation and imminent harm" rule do not comply with the requirements for restricting expression under Article 19(3). To bring the rule into alignment with the "provided by law" requirement, Facebook should provide a clear definition of relevant terms to provide clarity for users. Facebook should also elaborate on how they determine whether the threshold of "imminent harm" has been met. Facebook should always pursue the least restrictive means of content moderation to meet the policy goal.

CASE STUDY 4. KARABAKH IS AZERBAIJAN (KIAZ) : APPLYING INSTAGRAM PLATFORM LAW

FIGURE 4.1 KIAz's Facebook post from December 2, 2020 (translated using Google Translate).

Translation: Dear Rector of Baku Engineering University, Professor Havar Mammadov, and teachers, we thank the staff as a platform "Karabakh is Azerbaijan". Founded in June 2020 [sic] through online and social media platforms, our platform unites all differences and unites under one goal. This goal is to convey the realities of Azerbaijan and Karabakh to the world. Both our successful participation in the information war during the Second Karabakh War, as well as the successes we will achieve in our subsequent activities for the purpose, are the result of the value and assistance you provide, as well as our team. So, thanks to the conditions you have created for our volunteer team at Baku Engineering University, your support, access to the Internet, technical equipment, and being with us both materially and spiritually, we have successfully signed in the information war. Therefore, as the Karabakh is Azerbaijan platform, we sincerely thank all the university staff, including Professor Havar Mammadov.

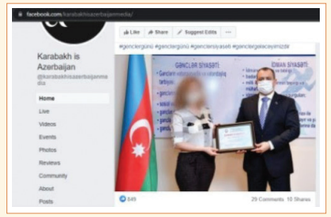

FIGURE 4.2 Image from Facebook of one of the co-founders of KIAz being handed an award by Azerbaijani Member of Parliament, Adil Aliyev.

The conflict witnessed the rise of student-driven patriotic astroturfing, which describes the phenomenon whereby small, coordinated accounts post at a high volume to create the impression that an online movement enjoys more support than it actually does.[221] A prominent example of a student-led activist group that was active during the conflict is the "Karabakh is Azerbaijan" (KIAz) platform, which operates Instagram,[222] Facebook,[223] Twitter,[224] YouTube,[225] and Telegram[226] accounts. As of August 3, 2021, the platform had 18.7k, 25k, 5.69k, 1.61k, and 601 followers on each of its accounts respectively.

According to KIAz's "About" page on YouTube, the platform was created by a group of "patriotic youth" on July 13, 2020, two days after the July 2020 clashes between Armenia and Azerbaijan had materialized.[227] Review of posts across their different platforms reveals that the group is supported by the Baku Engineering University, a university established in 2016 under the order of the President of Azerbaijan, Ilham Aliyev, to train students on how to become professionals in the field of engineering technology.[228] A Facebook post from KIAz (Figure 4.1) specifically thanks the rector of the university, Professor Havar Mammadov, who was appointed to this position under the order of President Aliyev, and goes on to express pride in the platform's contributions to the "information war" between Armenia and Azerbaijan that simultaneously took place during the conflict. According to their Facebook and Instagram accounts, the group was presented with two awards by an Azerbaijani Member of Parliament for their efforts during the "information war" (see Figures 4.2 and 4.3).

KIAz's commitment to the information war is evident across their social media accounts on different platforms. For example, on October 3, 2020, KIAz posted a video to their Instagram account. The video depicts two individuals, circled in red by the content creator, who appear not to be in military uniform and are helping with artillery preparations (see Figures

FIGURE 4.3 An image from Instagram of one of the co-founders of KIAz being granted an award by Azerbaijani Member of Parliament, Adil Aliyev.

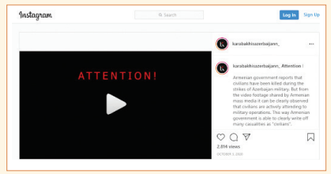

FIGURE 4.4 Screenshot of KIAz's Instagram post from October 3, 2020[229]

FIGURE 4.5 The men circled in red by the content creator are "civilians," according to KIAz.

FIGURE 4.6 Image of the real Hraparak TV's official Instagram account, which does not match the account that posted the video.

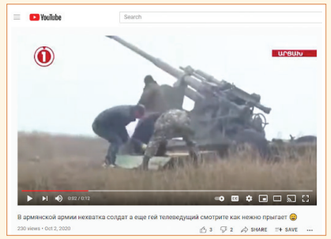

FIGURE 4.7 From YouTube.

4.4-4.5). The post's caption alleges that "from the video footage shared from Armenian mass media it can be clearly observed that civilians are actively attending to military operations. This way Armenian government is able to clearly write off many casualties as 'civilians.'"

There is a logo in the upper left corner of a screenshot taken from the video. Through a reverse image search on Yandex Images, we found an Instagram account run by a user named "hraparak_tv" that uses the same logo. The video was shared by that user on October 2, 2020, a day before the video was posted by KIAz, and contains a similar caption as the post made by KIAz. While there is a legitimate Armenian media outlet named Hraparak TV, this "harapak_tv" Instagram account is not an official account of that outlet (see Figure 4.6). The legitimate Hraparak TV has confirmed that the "hraparak_tv" Instagram account is an unofficial account run by Azerbaijanis, designed to look like the content being posted to that account is derived from legitimate Armenian media.[230]

Upon conducting further reverse image searches of the screenshots taken from the video using Yandex Images, we discovered that there were actually two versions of this video in circulation on social media—the one that was posted by both the unofficial "hraparak_tv" Instagram account and the KIAz Instagram account and a second video posted to YouTube (see Figure 4.7).[231]

Most notably, the logo in the video posted by KIAz does not match the logo seen in the screenshot of the YouTube video. We performed a reverse image search of the YouTube video logo on Yandex Images, which yielded several results, including a Tweet linking the original video featuring the same logo from the YouTube video.[232] The original video is from an Armenian media outlet, 1in.am, and was posted on October 1, 2020,[233] a day before the unofficial "hraparak_tv" posted the video.

In the center-left of the original YouTube video at Figure 4.7, there is a man who is hunched over and can be seen wearing a dark jacket, a pair of jeans, and striped shoes. This man's plain clothing seems to be the reason why KIAz claims that civilians are involved in Armenia's military

operations in its October 3, 2020 Instagram post as these are the elements that are circled in red by KIAz in Figure 3.5.[234] However, in reviewing the original video posted by 1in.am on October 1, 2020 in its entirety, it appears that the men were not civilians. Starting at the 2:45 minute mark, the reporter in the 1in.am video starts conversing with the men, whom the reporter refers to as "soldiers", asking them about the recent attacks in Nagorno-Karabakh/Artsakh and their work. This was confirmed by a member of our team who is fluent in the language being spoken. It is during this segment that the same man who KIAz alleges is a civilian in Figure 3.5 appears in the background. He is wearing the same jacket, jeans, and striped shoes that we previously identified, but underneath that jacket, he is wearing an army shirt, just like the rest of the soldiers in the video. This crucial part of the video was selectively edited out of the version that was posted by the unofficial "hraparak_tv," which was then reposted by KIAz on October 3, 2020. The unofficial "hraparak_tv" Instagram account also selectively edited the reporter's conversation with the soldiers out of the 1in.am video, which provides important context to show the men are not civilians. We found numerous instances where the Azerbaijani media made the same allegations based on this content.[235]

1. Compliance with Relevant Platform Standards

Instagram's Community Guidelines incorporate many of Facebook's Community Standards, including Facebook's "Violence and Incitement" policy.[236] This policy prohibits posting "[m]isinformation and unverifiable rumors that contribute to the risk of imminent violence or physical harm."[237] In an update to this policy, it was clarified that "[m]isinformation that does not put people at risk of imminent violence or physical harm but is rated false by third-party fact-checkers will be [downranked]" so that fewer people are exposed to it.[238] Additionally, Instagram has a policy on reducing the spread of false information, which, at the time of the conflict, was meager and provided an external link to a Facebook announcement about how Facebook addresses fake news. This Facebook announcement stated that technology and human review is used to identify and downrank false news that is rated false by third-party fact-checkers, provide more context on false news, or remove pages that repeatedly share false news.[239]

The KIAz's Instagram post appears to be misinformation and, therefore, raises a question under the "Violence and Incitement" policy as to whether it contributed to a risk of imminent violence or physical harm. According to the FBOB it is paramount to consider the local context and current situation in a State when assessing the "imminent harm" threshold.[240] Here, the post was made on October 3, 2020, just six days after the conflict began, when tensions between Armenia and Azerbaijan were already running high. By making false allegations about the Armenian government's military tactics, the post could have arguably spurred further enmity between the two sides, particularly when viewed in light of KIAz's stated goal of winning the information war for Azerbaijan. However, the post does not say anything derogatory against Armenians,[241] and it does not expressly or impliedly suggest a risk of imminent violence or physical harm against Armenia or Armenians. Consequently, despite containing false information about Armenians utilizing civilians in combat, it is unlikely that any content moderation action would be appropriate under the current rules.

2. Compliance with international human rights law

Misinformation and disinformation are not prohibited forms of expression under international human rights law. To respect human rights, content can only be restricted if the three cumulative criteria of Article 19(3) ICCPR are met. KIAz's Instagram post touches upon a matter of public interest, namely civilian involvement in the conflict, so it is categorized as political speech, which is subject to particularly strong protection under Article 19(2) ICCPR.[242] The relevant law in question for an Article 19(3) analysis is Facebook's "Violence and Incitement" policy and, to be exact, the "misinformation and imminent harm" rule contained in that policy, which also applies to Instagram. In its rationale for this policy, Facebook states that its goal is to prevent offline harm and threats to public safety,[243] Thus, the rule pursues the legitimate grounds of respecting the rights of others and protecting public safety and national security required by Article 19(3). However, as discussed above, the rule does not satisfy the "provided by law" requirement, as it is missing definitions of misinformation and imminent harm. The policy therefore fails to meet the requirements to justify a restriction of expression that respects human rights.

We conclude that Facebook and Instagram policies are inadequate to deal with content of this nature. Although KIAz is not a State-run platform, its immense popularity during the conflict suggests that it had the power to use its manipulative messaging to influence its tens of thousands of followers. Despite strong protections for freedom of expression, there may be policy reasons to seek to moderate content of this nature, particularly during an armed conflict in accordance with the position of the UN Working Group group on the issue of human rights and transnational corporations. Appropriate moderation for content like this might be affixing a label that warns users about the misinformation contained in the post or downranking the post so that fewer users could see it.

Removal of the single Instagram post under scrutiny here would probably not address the issue of the repeated manipulative messaging KIAz was sending as part of its astroturfing campaign to control the information war. In Facebook's False News announcement that Instagram linked to in its "reducing the spread of false news" policy at the time of the conflict, Facebook stated that it removes the pages of repeat offenders.[244] While account removals and suspensions should be a last-resort measure taken against misinformation, a temporary account suspension for KIAz's Instagram account may have been proportionate under Article 19(3) ICCPR in the circumstances that existed at the time.[245]

CONCLUSIONS

These posts represent a small sample of content that was posted to these four social media companies prior to, during, and after the Fall 2020 conflict. These examples illustrate the complexity of removal decisions facing social media companies and their digital content moderators, particularly when the nature of content is not immediately evident. Uncovering instances of manipulation and disinformation is time consuming work. This raises questions about what level of verification and authentication is practical for social media companies to carry out given the vast amount of content posted to platforms every minute of every day. While social media companies have attempted to develop platform law to guide them in their content moderation decision making, that platform law is often unclear or imprecise, and does not always meet the international human rights law threshold to permissibly restrict expression. All four social media companies have extensive provisions that prohibit hate speech on their platforms, which comports with the prohibitions found in Article 20(2) ICCPR. Yet, platforms are still failing to moderate hate speech. Each social media company needs to make an effort to more directly incorporate prohibitions on expression constituting propaganda for war (Article 20(1) ICCPR) into their platform law. Moreover, in many instances, platform law lacks the specificity required by Article 19(3) ICCPR to permit restrictions of expression on the platforms. In instances where platform law and international human rights law align, social media companies do not always apply their own policies successfully. In addition, despite strong protections for freedom of expression, there may be policy reasons to moderate content, particularly during an armed conflict where expression online may influence conflict dynamics on the ground.

Content removal is not something that should be undertaken lightly, given the right to freedom of expression and the concomitant risks associated with assessing and removing content. It is therefore especially important that social media companies develop robust platform law, guided by international human rights standards, to ensure that the rights of all platform users are upheld.

RECOMMENDATIONS

To aid social media companies in their policy development, and to support them in respecting human rights, we offer the following general recommendations:

- All social media companies should develop platform law that is guided by international human rights law, in particular Article 20 and Article 19 of the International Covenant on Civil and Political Rights (ICCPR).[246]
- All social media companies should ensure that their platform law is "formulated with sufficient precision to enable an individual to regulate his or her conduct accordingly" in accordance with the "provided by law" requirement of Article 19(3) ICCPR.[247]
- All social media companies should ensure that their platform law is "accessible to the public."[248] This may require social media companies to collate their moderation policies in a centralized location so that users can clearly see the rules.[249]
- Where social media companies provide for the possibility of content moderation, platform law should be explicit as to whether the restriction on expression is pursuing one of the permitted grounds under Article 19(3) (a) or (b).
- All social media companies should adopt a scale of content moderation mechanisms so that, where moderation is necessary under platform law and international human rights law, platforms are able to moderate the content using the least restrictive means possible.[250]
- Social media companies have a heightened duty to respect and protect human rights during armed conflict under the UN Guiding Principles on Business and Human Rights and the guidance provided by the UN Working Group on the issue of human rights and transnational corporations and other business enterprises.[251] The Working Group has identified three steps that social media companies should take, which should be implemented by all platforms:
  - (1) identify the root causes of tensions and potential triggers, including contextual factors, and the real and perceived grievances that are steering the conflict;[252]
  - (2) map the main actors in the conflict and their motives, capacities, and opportunities to inflict violence;[253] and
  - (3) identify and anticipate the ways in which the business' own operations, products, or services impact upon existing tensions and relationships between the various groups and/or create new tensions or conflicts.[254]

The "concrete steps that businesses need to take will be extremely context dependent."[255] It may include "suspend[ing] or terminat[ing] activities in or linked to a conflict-affected context,"[256] and social media companies should take steps to "anticipate and plan a clear exit strategy in advance."[257]

We also offer the following specific recommendations to individual platforms.

VK should make the following amendments and modifications to its "Platform Standards" policy:

- VK should update its policy to make explicit its definition of hate speech and clarify its evaluation criteria.
- VK should adopt more precise and detailed language to provide users with a greater appreciation for its values and how its policies might change in the future.
- VK should better enforce its existing policies, particularly during periods and in regions experiencing armed conflict.

Twitter should make the following amendments and clarifications to its "Hateful conduct policy":

- Twitter should clarify whether the "inciting fear about a protected category" rule applies if the content targets an entire group based on a protected category or only certain members of a group by providing more hypothetical examples as to the types of Tweets that would contravene its policy.
- Twitter should add an external link to its "Our approach to policy development and enforcement philosophy" page on Twitter's "Hateful conduct policy" so that users can easily find the factors Twitter considers when deciding what action to take against a piece of content.
- Twitter should clarify when it needs to hear from a person being targeted in a post in order to take action against a tweet.
- Twitter should explicitly make clear to users what falls within its protected categories by expressly saying that "Protected categories include…."
- Twitter should update its policy to include labeling as a potential consequence of violating this policy.
- Twitter should consider linking unbiased, trustworthy sources of information, when such sources are available, on Tweets that classify as propaganda for war to keep users informed of developing situations around an armed conflict.
- Twitter should make policy citations of accounts available to researchers while ensuring that any privacy-compromising information in those citations is kept to a minimum.

Twitter should make the following amendments and clarifications to its "Synthetic and manipulated media policy":

- Twitter's algorithms and third-party partners should prioritize synthetic and manipulated media shared in the context of an armed conflict.
- Twitter should make the human rights that it seeks to protect more explicit on the face of the policy's rationale in order to provide greater clarity.
- Twitter should add that links to trustworthy sources will be provided as part of its enforcement response if such trustworthy sources exist.
- Twitter should consider adopting a general policy on misinformation and disinformation or a policy on misinformation and disinformation in the context of an armed conflict. Whatever path Twitter chooses, it should set out clear definitions for misinformation and disinformation and list a range of enforcement options, including content removal as a last-resort measure where the content is likely to contribute to serious imminent harm.

Facebook and Instagram should make the following amendments and modifications to their "Violence and Incitement" policy:

- In its "misinformation and imminent harm" rule, Facebook should provide a clear definition of "misinformation," working in tandem with civil society organizations and other stakeholders to craft a suitable definition.
- In its "misinformation and imminent harm" rule, Facebook should elaborate on the signals it looks for in determining whether the "imminent harm" threshold has been met.
- In its "misinformation and imminent harm" rule, Facebook should make clear that repeat violations could result in temporary account suspensions.
- Facebook should provide greater transparency on how it partners with third-party fact-checkers and how its algorithms and human review processes work.
- Facebook should prioritize referring content that makes dubious claims about an armed conflict to its fact-checkers.
- Facebook should expand its policy to include alternative measures to downranking, such as affixing a label that warns users of the misinformation and/or directs users to trusted sources of information.

ENDNOTES

[1] *See* Christopher Giles & Upasana Bhat, *Nagorno-Karabakh: The Armenian-Azeri 'information wars,'* BBC (Oct. 26, 2020), https://www.bbc.com/news/world-europe-54614392. The Oxford Internet Institute concluded that the governments of Armenia and Azerbaijan, or private firms they work with, employed cyber troop teams consisting of full-time staff members coordinating with multiple actors to control the information space, including to advance disinformation campaigns. *See* Samantha Bradshaw et al., *Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation*, 18-19 (Jan. 13, 2021), https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/127/2021/01/CyberTroop-Report-2020-v.2.pdf.

[2] *See e.g.* International Committee of the Red Cross (ICRC), *Harmful Information - Misinformation, disinformation and hate speech in armed conflict and other situations of violence: ICRC initial findings and perspectives on adapting protection approaches*, 5 (July 9, 2021), https://shop.icrc.org/harmful-information-misinformation-disinformation-and-hate-speech-in-armed-conflict-and-other-situations-of-violence-icrc-initial-findings-and-perspectives-on-adapting-protection-approaches-pdf-en [hereinafter ICRC report] at 11.

[3] This is enshrined in principle 11 of the UN Guiding Principles on Business and Human Rights ("UNGPs"), adopted by the Human Rights Council in 2011. John Ruggie (Special Representative of the Secretary-General), *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*, UN Doc. A/HRC/17/31, principle 11 (Mar. 21, 2011) [hereinafter UN Guiding Principles]. David Kaye, the former UN Special Rapporteur on freedom of opinion and expression, has been at the vanguard of advocating for the application of the UNGPs' "responsibility to respect" framework as a launchpad for reforming social media companies' platform policies. Kaye champions "smart regulation," whereby States focus on bolstering company transparency while social media companies engage in a "human rights by default" approach with respect to their platforms. David Kaye, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, UN Doc A/HRC/38/35, ¶ 1 (Apr. 6, 2018) [hereinafter Kaye 2018 Report].

[4] Working Group on the issue of human rights and transnational corporations and other business enterprises, *Human rights and conflict-affected regions: towards heightened action*, UN Doc A/75/212, ¶¶ 13 (July 21, 2020) [hereinafter Working Group Report].

[5] Kaye 2018 Report, *supra* note 3.

[6] ICRC report, *supra* note 1, at 11.

[7] International Covenant on Civil and Political Rights, art.19(2), *opened for signature* Dec. 16, 1966 (entered into force Mar. 23, 1976) [hereinafter ICCPR].

[8] Irene Khan, *Disinformation and freedom of opinion and expression*, UN Doc A/HRC/47/25, ¶ 38 (Apr. 13, 2021) [hereinafter Khan 2021 Report].

[9] ICCPR, *General comment No. 34, Article 19: Freedoms of opinion and expression*, UN Doc CCPR/C/GC/34, ¶ 12 (Sept. 12, 2011) [hereinafter GC 34].

[10] *Id.* ¶ 50.

[11] Khan 2021 Report, *supra* note 8 ¶ 39.

[12] *Id.* at ¶ 38.

[13] *Id*.

[14] Kaye 2018 Report, *supra* note 3.

[15] Emma Llansó et al., *Artificial Intelligence, Content Moderation, and Freedom of Expression,* Transatlantic Working Group, (Feb. 26, 2020), https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf.

[16] *See e.g.* Giovanni De Gregorio, *Democratising online content moderation: A constitutional framework*, 2020, https://doi.org/10.1016/j.clsr.2019.105374; Sahana Udupa et al., *Artificial Intelligence, Extreme Speech, and the Challenges of Online Content Moderation,* 2021, AI4Dignity

[17] UN Guiding Principles, *supra* note 3.

[18] *See e.g.*, Susan Benesch, *But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies*, 38:86 Yale J. on Reg. Bulletin 86, 92 (2020).

[19] UN Guiding Principle 11: "Business enterprises should respect human rights. This means that they should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved." According to principle 12, this responsibility entails that businesses should, at a minimum, respect the International Bill of Rights, consisting of the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights ("ICCPR"), and the International Covenant on Economic, Social and Cultural Rights. UN Guiding Principles *supra* note 3.

[20] Working Group Report, *supra* note 4, at ¶¶ 13, 43. The Working Group identified the following list of factors as relevant in enabling businesses to recognize when they are subject to this heightened responsibility: the existence of an armed conflict or other forms of instability between States; weakness or absence of State structures; a State's record of serious violations of international human rights law and IHL; the amassing of weapons or arms; the imposition of emergency laws or other security measures; the suspension of, or interference with, vital State institutions, particularly if this results in the exclusion of vulnerable or minority groups; increased politicization of identity; increased inflammatory rhetoric or hate speech; sustained signs of militia or paramilitary groups; the strengthening of security apparatuses or mobilization against specific groups by States; strict control or banning of communication channels; and the expulsion or banning of civil society organizations, the media, and any other watchdogs. *Id.* at ¶¶ 16-21.

[21] *Id*. at ¶ 98.

[23] UN Guiding Principles, *supra* note 3.

[23] ICCPR, *supra* note 7, at art.19(2).

[24] Khan 2021 Report, *supra* note 8, at ¶ 38.

[25] GC 34, *supra* note 9, at ¶ 12.

[26] *Id.* at ¶ 22.

[27] ICCPR, *supra* note 7, at art.20(1). The terms "propaganda for war" and "war propaganda" have been used interchangeably. *See*, *e.g.*, OSCE Office of the Representative on Freedom of the Media, *Propaganda and Freedom of the Media*, 17 (2015), https://www.osce.org/files/f/documents/b/3/203926.pdf. "Propaganda for war" seems to apply only to expressions made *prior* to the outbreak of armed conflict. "War propaganda" is understood as a broad tool that is typically weaponized *during* an armed conflict to promote the national war effort domestically, to demoralize the enemy, to strengthen relations among allies, or to compel neutral States to join one side of the war or another. *See* Ralph D. Casey, *What Is Propaganda?*, American Hist. Ass'n (July 1944), https://www.historians.org/about-aha-and-membership/aha-history-and-archives/gi-roundtable-series/pamphlets/em-2-what-is-propaganda-(1944)/war-propaganda.

[28] Michael G. Kearney, *The Prohibition of Propaganda for War in International Law* 132 (2007) [hereinafter Kearney].

[29] In Resolution 381(V), the General Assembly declared that propaganda includes (1) incitement to conflicts or acts of aggression, (2) measures tending to isolate the peoples from any contact with the outside world, and (3) measures tending to silence or distort the UN's activities in favor of peace or to prevent their peoples from knowing the views of other Member States. *See* GA Res 381(V), UNGA, 5th Sess, UN Doc A/RES/381(V) (Nov. 17, 1950). *See also* Resolution 110(II), in which the UN General Assembly condemned all forms of propaganda that are "designed or likely to provoke or encourage any threat to the peace, breach of the peace, or act of aggression…." G.A. Res. 110(II), UNGA, 2d Sess, UN Doc A/RES/2/110 (Nov. 3, 1947).

[30] Third Committee, UNGA, 16th Sess, UN Doc. A/C.3/SR.1079, ¶ 2 (Mr. Mello) (Oct. 20, 1961) [hereinafter Third Committee proceedings]. *See also* Kearney, *supra* note 28, at 132.

[31] Kaye 2018 Report, *supra* note 3.

[32] ICCPR, *supra* note 7 at art.20(2). The International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) contains a corresponding provision to Art.20(2) ICCPR in Article 4, but this report exclusively discusses the ICCPR because the ICCPR's hate speech provision is broader, encompassing incitement to hostility or violence, not just racial discrimination. With that being said, most, if not all, of the relevant considerations that apply to the relationship between Articles 19 and 20(2) ICCPR also apply with equal force to the relationship between Articles 19 ICCPR and Art.4 ICERD. *See* International Convention on the Elimination of All Forms of Racial Discrimination, art.4, *opened for signature* Dec. 21, 1965 (entered into force Jan. 4, 1969).

[33] UN Strategy and Plan of Action on Hate Speech, 2 (June 18, 2019), https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf [hereinafter UN Hate Speech Strategy 2019].

[34] UNHCHR, Rabat Plan of Action, UN Doc. A/HRC/22/17/Add.4, ¶ 29 (Jan. 11, 2013) [hereinafter Rabat Plan]. This test was originally adopted to assess expression that should be criminalized, but has since been expanded. Essentially, the more severe the hate speech in question is under the Rabat test, the more likely that restricting the hate speech will be necessary and proportionate under the third prong of the test in Article 19(3) ICCPR, *supra* note 7.

[35] Another view, propounded by the OSCE Representative on Freedom of the Media, is that "freedom of expression under the ICCPR should be interpreted as *not including* war propaganda and hate speech…. According to this view, restrictions on "propaganda for war" or "hate speech" would not need to satisfy the tripartite test in Article 19(3). *See* OSCE 2015 Report, *supra* note 28. This parallels how the European Court of Human Rights (ECtHR) sometimes treats hate speech. *See e.g.* Ethan Shattock, *Should the ECtHR Invoke Article 17 for Disinformation Cases?*, EJIL:Talk! (Mar. 26, 2021), https://www.ejiltalk.org/should-the-ecthr-invoke-article-17-for-disinformation-cases/.

[36] *See e.g.* Khan 2021 Report, *supra* note 8; *see e.g.* European Parliament, *The impact of disinformation on democratic processes and human rights in the world*, (Apr. 2021), https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf

[37] *See e.g.* Khan 2021 Report, *supra* note 8, at ¶ 1; The Digital, Culture, Media, and Sport Committee, *Disinformation and 'fake news': Interim Report: Government Response to the Committee's Fifth Report of Session 2017-19*, 2 (Oct. 23, 2018), https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1630/1630.pdf; Veridiana Alimonti, *EFF to the inter-American System: If You Want to Tackle "Fake News," Consider Free Expression First*, EFF (Feb. 28, 2019), https://www.eff.org/deeplinks/2019/02/eff-inter-american-system-if-you-want-tackle-fake-news-think-free-expression-first; Darrin Baines & Robert J.R. Elliott, *Defining misinformation, disinformation and malinformation: An urgent need for clarity during COVID-19 infodemic*, 16 (Apr. 21, 2020), http://www.repec.bham.ac.uk/pdf/20-06.pdf; Joshua A. Tucker et al., *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature*, 55 (Mar. 2018), https://www.hewlett.org/wp-content/uploads/2018/03/Social-Media-Political-Polarization-and-Political-Disinformation-Literature-Review.pdf.

[38] Khan 2021 Report, *supra* note 8, at ¶ 38.

[39] Statement by Irene Khan, Special Rapporteur on the promotion and protection of the freedom of opinion and expression to the 47th Session of the Human Rights Council, Jul. 2, 2021, https://www.ohchr.org/en/press-briefing-notes/2021/07/statement-irene-khan-special-rapporteur-promotion-and-protection.

[40] *See, e.g.*, *Our approach to policy development and enforcement philosophy*, Twitter Help Center, https://help.twitter.com/en/rules-and-policies/enforcement-philosophy [hereinafter Twitter enforcement philosophy]; *Community Standards*, Facebook, https://www.facebook.com/communitystandards/ [hereinafter Facebook Community Standards]; *Community Guidelines*, Instagram Help Center, https://help.instagram.com/477434105621119 [hereinafter Instagram Community Guidelines]; *Safety Guidelines*, VK, https://m.vk.com/safety?section=social&lang=en [hereinafter VK Safety Guidelines].

[41] Case Decision 2020-006-FB-FBR (Facebook Oversight Board 2021), https://www.oversightboard.com/decision/FB-XWJQBU9A. If not listed under their Community Standards or Community Guidelines respectively, then Facebook and Instagram's policies can be found on Facebook's Newsroom.

[42] It should be noted that some platforms do not make their policy changes easily searchable and so in some instances we were unable to find information regarding modifications that may have been made between November 2020 and August 2021. We were able to track some changes to the platform policies using the Wayback Machine (https://archive.org/), archive.today (https://archive.ph/), and Letter Girl (https://letrachica.digital/). However, The Wayback Machine cannot archive Facebook and Instagram sites, and VK's platform policies have not been archived there.

[43] *Terms of Service*, VK, ¶ 8.5, (last updated May 21, 2018), https://vk.com/terms [hereinafter VK Terms of Service]. *See also* VK Safety Guidelines, *supra* note 40.

[44] VK Safety Guidelines, *supra* note 40.

[45] *See Platform Standards*, VK, https://m.vk.com/safety?lang=en&section=standards.

[46] *Id.*

[47] *Id.*

[48] VK *Terms of Service*, *supra* note 44

[49] Twitter enforcement philosophy, *supra* note 43.

[50] *Id.*

[51] *The Twitter Rules*, Twitter Help Center, https://help.twitter.com/en/rules-and-policies/twitter-rules [hereinafter Twitter Rules].

[52] *Id.*

[53] *Violent threats policy*, Twitter Help Center (Mar. 2019), https://help.twitter.com/en/rules-and-policies/violent-threats-glorification.

[54] *Glorification of violence policy*, Twitter Help Center (Mar. 2019), https://help.twitter.com/en/rules-and-policies/glorification-of-violence.

[55] *Abusive behavior policy*, Twitter Help Center, https://help.twitter.com/en/rules-and-policies/abusive-behavior.

[56] *Hateful conduct policy*, Twitter Help Center, https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

[57] *Sensitive media policy*, Twitter Help Center (Nov. 2019), https://help.twitter.com/en/rules-and-policies/media-policy.

[58] Twitter Rules, *supra* note 52.

[59] *Platform manipulation and spam policy*, Twitter Help Center (Sept. 2020), https://help.twitter.com/en/rules-and-policies/platform-manipulation.

[60] *Impersonation policy*, Twitter Help Center, https://help.twitter.com/en/rules-and-policies/twitter-impersonation-policy.

[61] *Synthetic and manipulated media policy*, Twitter Help Center, https://help.twitter.com/en/rules-and-policies/manipulated-media.

[62] Twitter enforcement philosophy, *supra* note 40.

[63] *Id.*

[64] *Our range of enforcement options,* Twitter Help Center, https://help.twitter.com/en/rules-and-policies/enforcement-options

[65] *Id.*

[66] *Corporate Human Rights Policy*, Facebook (Mar. 16, 2021), https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf [hereinafter Facebook Human Rights Policy].

[67] *Id.*

[68] *Facebook Community Standards*, *supra* note 41.

[69] *Id.*

[70] *Id.*

[71] *Violence and Incitement*, Facebook, https://www.facebook.com/communitystandards/credible_violence [hereinafter Facebook Violence and Incitement].

[72] *Dangerous Individuals and Organizations*, Facebook, https://www.facebook.com/communitystandards/dangerous_individuals_organizations [hereinafter Facebook Dangerous Individuals and Organizations Policy].

[73] *Coordinating Harm and Publicizing Crime*, Facebook, https://www.facebook.com/communitystandards/coordinating_harm_publicizing_crime.

[74] Facebook Dangerous Individuals and Organizations Policy, *supra* note 72.

[75] Facebook Violence and Incitement, *supra* note 71.

[76] *Hate Speech*, Facebook, https://www.facebook.com/communitystandards/hate_speech.

[77] *Violent and Graphic Content*, Facebook, https://www.facebook.com/communitystandards/graphic_violence.

[78] Facebook Community Standards, *supra* note 40.

[79] *Account Integrity and Authentic Identity*, Facebook, https://www.facebook.com/communitystandards/misrepresentation.

[80] *False News*, Facebook, https://www.facebook.com/communitystandards/false_news.

[81] *Manipulated Media*, Facebook, https://www.facebook.com/communitystandards/manipulated_media.

[82] Sam Shead, *Facebook owns the four most downloaded apps of the decade*, BBC (Dec. 18, 2019), https://www.bbc.com/news/technology-50838013.

[83] Facebook Human Rights Policy, *supra* note 66.

[84] Instagram Community Guidelines, *supra* note 40.

[85] *Reducing the Spread of False Information on Instagram*, Instagram Help Center, https://help.instagram.com/1735798276553028.

[86] Third Committee proceedings, *supra* note 31, ¶ 2. *See also* Kearney, *supra* note 28, at 132.

[87] Khan 2021 Report, *supra* note 8, at ¶ 38

[88] GC 34, *supra* note 9, at ¶ 22.

[89] *Id.* at ¶ 25.

[90] *See* Kaye 2018 Report, *supra* note 3, at ¶ 46.

[91] GC 34, *supra* note 9, at ¶ 22.

[92] *Id.*, at ¶¶ 33-34.

[93] Working Group Report, *supra* note 4, at ¶¶ 46-48.

[94] Neither side accepts responsibility for striking first. HRW wrote that "September 27, Azerbaijan launched a military offensive that escalated hostilities between Azerbaijan and Armenia and the de-facto authorities in Nagorno-Karabakh." *Azerbaijan Events of 2020*, HRW, https://www.hrw.org/world-report/2021/country-chapters/azerbaijan.

[95] Todd Carney, *Applying International Law to the Nagorno-Karabakh Conflict*, Opinio Juris (Jan. 1, 2020), http://opiniojuris.org/2020/01/22/applying-international-law-to-the-nagorno-karabakh-conflict/.

[96] Beginning in 1988 there was high tension and later guerilla warfare. Then in 1992, the full-scale war began. *See Nagorno Karabakh Conflict*, Council on Foreign Relations, https://www.cfr.org/global-conflict-tracker/conflict/nagorno-karabakh-conflict. By the time the war concluded with a Ceasefire Agreement in 1994, Armenia had full control of the land that was historically Artsakh, as well as some of the surrounding Azerbaijani territories. *See The Nagorno-Karabakh Conflict: A Visual Explainer*, Crisis Group, https://www.crisisgroup.org/content/nagorno-karabakh-conflict-visual-explainer. Yet, negotiations run by the OSCE Minsk Group have failed to establish a Peace Treaty. The OSCE Minsk Group "spearheads the OSCE's efforts to find a peaceful solution to the Nagorno-Karabakh conflict." France, Russia, and the United States are co-chairs. *See OSCE Minsk Group*, OSCE, https://www.osce.org/mg. Azerbaijan continues to contend that all of the territory gained during the war is Azerbaijani but under Armenian occupation. *See* Gerard Toal and John O'Loughlin, *Here are the 5 things you need to know about the deadly fighting in Nagorno Karabakh*, Wash. Post (Apr. 16, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/04/06/will-war-erupt-in-nagorny-karabakh-here-are-the-5-things-you-need-to-know/. Since the 1994 Ceasefire, there have been constant skirmishes along the Line of Contact, including the Four-Day War in 2016 and a conflict within the territory of Armenia along the Armenian/Azerbaijani Border in July 2020. *See e.g.*, *Azerbaijan: Seven Years of Conflict in Nagorno-Karabakh*, HRW (Dec. 1, 1994), https://www.hrw.org/report/1994/12/01/seven-years-conflict-nagorno-karabakh. Avet Demourian, *Armenia-Azerbaijan border fighting escalates; 16 killed*, Wash. Post (July 15, 2020), https://www.washingtonpost.com/world/europe/correction-armenia-azerbaijan-story/2020/07/15/d1606484-c6d7-11ea-a825-8722004e4150_story.html.

[97] *See e.g., Azerbaijan: Unlawful Strikes in Nagorno-Karabakh*, HRW (Dec. 11, 2020), https://www.hrw.org/news/2020/12/11/azerbaijan-unlawful-strikes-nagorno-karabakh; *Armenia: Unlawful Rocket, Missile Strikes on Azerbaijan*, HRW (Dec. 11, 2020), https://www.hrw.org/news/2020/12/11/armenia-unlawful-rocket-missile-strikes-azerbaijan; *Azerbaijan: Attack on Church Possible War Crime*, HRW (Dec. 16, 2020), https://www.hrw.org/news/2020/12/16/azerbaijan-attack-church-possible-war-crime; *Azerbaijan/Armenia: Scores of civilians killed by indiscriminate use of weapons in conflict over Nagorno-Karabakh*, Amnesty Int'l (Jan. 14, 2021), https://www.amnesty.org/en/latest/news/2021/01/azerbaijan-armenia-scores-of-civilians-killed-by-indiscriminate-use-of-weapons-in-conflict-over-nagorno-karabakh/;

[98] *See, e.g., Hugh Williamson, Unlawful Attacks on Medical Facilities and Personnel in Nagorno-Karabakh,* HRW (Feb. 26, 2021), https://www.hrw.org/news/2021/02/26/unlawful-attacks-medical-facilities-and-personnel-nagorno-karabakh.

[99] *See, e.g.*, Promise Institute for Human Rights, *Nagorno-Karabakh Conflict: Introduction*, https://libguides.law.ucla.edu/NagornoKarabakhConflict.

[100] *See e.g., Azerbaijan: Cluster Munitions Used in Nagorno-Karabakh*, (Oct. 23, 2020), https://www.hrw.org/news/2020/10/23/azerbaijan-cluster-munitions-used-nagorno-karabakh; *Armenia: Cluster Munitions Used in Multiple Attacks on Azerbaijan*, (Dec. 15, 2020), https://www.hrw.org/news/2020/12/15/armenia-cluster-munitions-used-multiple-attacks-azerbaijan; Azerbaijan: It was also alleged that Azerbaijan used white phosphorus, a banned substance. See e.g. *Satellite imagery shows environmental damage of reported white phosphorus use in Nagorno Karabakh*, Atlantic Council Digital Forensic Research Lab, (Nov. 12, 2020), https://medium.com/dfrlab/satellite-imagery-shows-environmental-damage-of-reported-white-phosphorus-use-in-nagorno-karabakh-9826391a295.

[101] *See e.g.* Azerbaijan: Armenian Prisoners of War Badly Mistreated, HRW (Dec. 2, 2020), https://www.hrw.org/news/2020/12/02/azerbaijan-armenian-prisoners-war-badly-mistreated; *Survivors of unlawful detention in Nagorno-Karabakh speak out about war crimes,* HRW (Mar. 12, 2021), https://www.hrw.org/news/2021/03/12/survivors-unlawful-detention-nagorno-karabakh-speak-out-about-war-crime; *Azerbaijan: Armenian POW's Abused in Custody,* HRW (Mar. 19, 2021), *https://www.hrw.org/news/2021/03/19/azerbaijan-armenian-pows-abused-custody.*

[102] *Armenia, Azerbaijan and Russia sign Nagorno-Karabakh peace dea*l, BBC (Nov. 10, 2020), https://www.bbc.com/news/world-europe-54882564.

[103] *See e.g. Freedom on the Net 2021: Azerbaijan,* Freedom House, https://freedomhouse.org/country/azerbaijan/freedom-net/2021; Katy Pearce, *While Armenia and Azerbaijan fought over Nagorno-Karabakh, their citizens battled on social media*, Wash. Post (Dec. 4, 2020, 7:45 AM), https://www.washingtonpost.com/politics/2020/12/04/while-armenia-azerbaijan-fought-over-nagorno-karabakh-their-citizens-battled-social-media/

[104] *Id.*

[105] *Media and disinformation in the Nagorno-Karabakh conflict and their role in conflict resolution and peacebuilding*, College of Europe, 9 (Jan. 2021), at p 9, https://www2.coleurope.eu/system/tdf/uploads/news/event_report_-_media_and_disinformation_in_the_nagorno-karabakh_conflict.pdf?&file=1&type=node&id=draft&force= [hereinafter College of Europe Report].

[106] *Id.*

[107] *October 2020 Coordinated Inauthentic Behavior Report*, About Facebook (Oct. 2020), https://about.fb.com/wp-content/uploads/2020/11/October-2020-CIB-Report.pdf. Coordinated inauthentic behavior (CIB) consists of "coordinated efforts to manipulate public debate for a strategic goal where fake accounts are central to the operation." CIB can occur in the context of domestic, non-government campaigns or on behalf of a foreign or government actor. The accounts and pages that were removed by Facebook were all linked to the Youth Union of New Azerbaijani Party.

[108] *See, e.g.*, Sabina Garahan, *False Equivalences in the Nagorno-Karabakh Conflict — International Humanitarian and Criminal Law Perspectives*, Opinio Juris (Feb. 2, 2021), http://opiniojuris.org/2021/02/10/false-equivalences-in-the-nagorno-karabakh-conflict-international-humanitarian-and-criminal-law-perspectives/; Carlotta Gall, *In Azerbaijan, a String of Explosions, Screams and Then Blood*, The N.Y. Times (Oct. 29, 2020), https://www.nytimes.com/2020/10/28/world/europe/azerbaijan-barda-armenia-rockets-karabakh.html; Shushan Stepanyan (@ShStepanyan), TWITTER (Oct. 27, 2020, 10:33 AM), https://twitter.com/ShStepanyan/status/1321097672711950337 [https://web.archive.org/web/20201027143534/https://twitter.com/ShStepanyan/status/1321097672711950337].

[109] Arshaluys Barseghyan et al., *Disinformation and Misinformation in Armenia: Confronting the Power of False Narratives*, Freedom House, 18 (June 2021), https://freedomhouse.org/sites/default/files/2021-06/Disinformation-in-Armenia_En-v3.pdf.

[110] *Id.* at p 18.

[111] College of Europe Report, *supra* note 106.

[112] *See*, *e.g.*, Twitter (Oct. 25, 2020, 11:16 AM), https://twitter.com/Gunel99M/status/1320383678011314176 [https://web.archive.org/web/20201027054507/https://twitter.com/Gunel99M/status/1320383678011314176]; INSTAGRAM (Oct. 28, 2020, 4:31 AM), https://www.instagram.com/p/CG4W9pVFGT0/; Hikmet Hajiyev (@HikmetHajiyev), Twitter (Oct. 24, 2020, 2:01 PM), https://mobile.twitter.com/HikmetHajiyev/status/1320062999453929474 [https://web.archive.org/web/20201024180400/https://twitter.com/HikmetHajiyev/status/1320062999453929474].

[113] *See*, *e.g.*, Twitter (Oct. 2, 2020, 7:57 PM), https://twitter.com/mert___3434/status/1312179941060956161; Ali Alizada (@Ali_F_Alizada), Twitter (Oct. 2, 2020, 1:54 AM), https://twitter.com/Ali_F_Alizada/status/1311907438434619394; https://www.instagram.com/p/CFz3ATel1iQ/ [https://web.archive.org/web/20201007061159/https://twitter.com/Ali_F_Alizada/status/1311907438434619394]. It has been reported that Syrian mercenaries were deployed to Azerbaijan, while claims about foreign mercenaries in Armenia either remain contested or have been debunked. See e.g. Bethan McKernan, *Syrian rebel fighters prepare to deploy to Azerbaijan in sign of Turkey's ambition*, The Guardian (Sept. 28, 2020, 2:13 PM), https://www.theguardian.com/world/2020/sep/28/syrian-rebel-fighters-prepare-to-deploy-to-azerbaijan-in-sign-of-turkeys-ambition; Karine Ghazaryan, *Wagner-Affiliated Telegram Channel Trolls Nagorno-Karabakh Conflict Analysts*, Bellingcat (Oct. 7, 2020), https://www.bellingcat.com/news/uk-and-europe/2020/10/07/wagner-affiliated-channel-trolls-nagorno-karabakh-conflict-analysts/.

[114] *Freedom on the Net 2021: Azerbaijan*, Freedom House, https://freedomhouse.org/country/azerbaijan/freedom-net/2021;

[115] *See*, *e.g.*, Hikmet Hajiyev (@HikmetHajiyev), Twitter (Feb. 17, 2020, 12:53 PM), https://twitter.com/HikmetHajiyev/status/1229463798853050370?s=20 [https://web.archive.org/web/20200217175515/https://twitter.com/HikmetHajiyev/status/1229463798853050370]. *See*, *e.g.*, Case decision 2020-003-FB-UA (Facebook Oversight Board 2021), https://oversightboard.com/decision/FB-QBJDASCV/.

[116] This process was conducted in early 2020, at which point we identified 40 hashtags and an additional 30 search terms, events and dates that were commonly being used across different platforms.

[117] It should be noted that, in recent years, Facebook has implemented significant modifications to its platform in an effort to clamp down on coordinated behavior and inauthentic accounts, rendering it extremely difficult for researchers to conduct investigations on the platform. Our team encountered numerous problems in using Facebook during our research, which prevented us from gauging the full scope of the type of content that was circulating prior to and during the conflict on that platform.

[118] *See, e.g.*, Case decision 2021-008-FB-FBR (Facebook Oversight Board 2021), https://oversightboard.com/decision/FB-B6NGYREK/ [hereinafter Oversight Board Brazil COVID-19 decision].

[119] *Interview (Exclusive)*, ORDU.AZ (July 9, 2019, 12:58 AM), https://ordu.az/az/news/151791/radius-artiq-unikal-ve-tekrari-olmayan-bir-mehsul-kimi-efirde-oz-yerini-berkidib-musahibe-ekskluziv.

[120] Manana Hakobyan et al., *2020 Armenia-Azerbaijan Twitter War: An Investigation of Patriotic Astroturfing during the 2020 Armenia-Azerbaijan War*, DataPoint Armenia 1, 33-34 (2021), https://datapoint.am/dziv/, [hereinafter Astroturfing Report].

[121] Hans Kloss is the code name of the Soviet agent protagonist in the World War II-set Polish television series "Stawka większa niż życie". Additionally, Zahid Abdulov, a victim of the Khojaly killings, was known by the nickname "Hans Kloss", which may have served as inspiration for its use here or is a reflection of the popularity of the Polish show in Azerbaijan. *See* IMDB, Stawka większa niż życie, https://www.imdb.com/title/tt0065035/?ref_=nm_knf_t2; Sariyya Muslumgizi, Guest from Khojaly 249-250 (Adila Agabeyli & Taleh Bulud trans., Mammad Nazimoghlu & Asly Khalilgizi eds., 2008).

[122] Polygon Azerbaijan, VK, https://vk.com/polygon_az. [http://web.archive.org/web/20210925204055/https://vk.com/polygon_az].

[123] PLGN Azerbaijan, YouTube, https://www.youtube.com/channel/UCKDgCSbPv6Gg4V0e290MGSg. [http://web.archive.org/web/20210925204308/https://www.youtube.com/channel/UCKDgCSbPv6Gg4V0e290MGSg].

[124] Polygon Azerbaijan, Facebook, https://www.facebook.com/polygon.azerbaijan.

[125] World Intellectual Property Organization (WIPO), https://www3.wipo.int/branddb/en/.

[126] *See* Heydar Mirza, LinkedIn, https://az.linkedin.com/in/heydar-mirza-552a1088.; *see also* Jeffrey Mankoff, *Iran-Azerbaijan Relations and Strategic Competition in the Caucasus*, CSIS Presents (Apr.il 29, 2013), https://www.csis.org/events/iran-azerbaijan-relations-and-strategic-competition-caucasus.

[127] Caliber.az, https://caliber.az. [http://web.archive.org/web/20210925204753/https://caliber.az/]; Caliber.az, YouTube, https://www.youtube.com/c/CaliberAz, [http://web.archive.org/web/20201112223356/https://www.youtube.com/c/CaliberAz].

[128] RADIUS, Facebook, https://www.facebook.com/itvradius; RADIUS, YouTube, https://www.youtube.com/playlist?list=PLmBIdPJozgD4GTgQUKFyjUG6lwA_9_F99, [http://web.archive.org/web/20210925204957/https://www.youtube.com/playlist?list=PLmBIdPJozgD4GTgQUKFyjUG6lwA_9_F99].

[129] Referring to Armenians as "vermin" or a species of vermin was and is a theme in Azerbaijani online rhetoric and has even been used by such official sources as the President of Azerbaijan. *See* Joe Nerssessian, *The Mixed Messaging of Ilham Aliyev*, EVN Report (Oct. 22, 2020), https://www.evnreport.com/politics/the-mixed-messaging-of-ilham-aliyev.

[130] Google Translate, https://translate.google.com/.

[131] Yandex Translate, https://translate.yandex.com/.

[132] Polygon Azerbaijan, VK (July 2, 2020, 9:17 PM), https://vk.com/wall-66054882_46499 [http://web.archive.org/web/20210925205332/https://vk.com/wall-66054882_46499].

[133] Polygon Azerbaijan, VK (Oct. 11, 2020, 5:19 AM), https://vk.com/wall-66054882_56010, [http://web.archive.org/web/20210916163036/https://vk.com/wall-66054882_56010].

[134] Polygon Azerbaijan, VK (Oct. 18, 2020, 11:10 PM), https://vk.com/wall-66054882_56851, [http://web.archive.org/web/20210925210430/https://vk.com/wall-66054882_56851].

[135] The Azerbaijani government claimed military breakthroughs in the Fuzuli region on the night of October 17, 2020 and many resulting Armenian casualties; *see* Defense Ministry: Azerbaijan Army troops managed to move forward in various directions of front, AZERTAC, (Oct. 17, 2020, 11:59 PM), https://azertag.az/en/xeber/Defense_Ministry_Azerbaijan_Army_troops_managed_to_move_forward_in_various_directions_of_front-1616417.

[136] Margarita Achikyan, *Armenian Noses*, YouTube (Dec. 22, 2014), https://www.youtube.com/watch?v=Nb42OAtBpM8; Emil Babayan, *Everybody nose: the Armenian feature you simply can't avoid*, The Calvert Journal, (Jan. 25, 2016), https://www.calvertjournal.com/articles/show/5321/armenian-nose-national-symbol-rhinoplasty-yerevan.

[137] Don Harrán, *The Jewish nose in early modern art and music*, 28 Renaissance Stud. 50, 50 (2014) https://doi.org/10.1111/rest.12006.

[138] Polygon Azerbaijan, VK (Dec. 29, 2020, 1:20 AM), https://vk.com/wall-66054882_59773 [http://web.archive.org/web/20210925211548/https://vk.com/wall-66054882_59773].

[139] Gagik Shamshyan, *Tragic car accident in Syunik region □ Mitsubishi crashed into the abyss about 400 meters, appearing in the Varkarn River □ The dead are employees of the RA Ministry of Defense □ PHOTO REPORT*: Shamshyan.com (Dec. 29, 2020 9:56 AM), https://shamshyan.com/hy/article/2020/12/29/1174823/.

[140] Polygon Azerbaijan, VK (Mar. 23, 2021, 4:15 AM), https://vk.com/wall-66054882_61313 [http://web.archive.org/web/20210925213710/https://vk.com/wall-66054882_61313].

[141] *One of the missing servicemen found dead - MoD*, Panorama.am (Mar. 23, 2021, 1:38 AM), https://www.panorama.am/en/news/2021/03/23/missing-servicemen-Jermuk/2473604.

[142] The Battle of Zangezeur (1918-1921) is remembered as a critical victory over Soviet forces which preserved Armenian cultural and ethnic heritage. See Tsovinar Petrosyan, *Zangezur – The Battle for the Right to Remain Armenians – 1918-1921*, Art-A-Tsolum (Apr. 21, 2019), https://allinnet.info/history/zangezur-the-battle/.

[143] *See, e.g.*, Twitter, https://twitter.com/ScourgeOfTengri/status/1397288588191178753, [http://web.archive.org/web/20210925214033/https://twitter.com/ScourgeOfTengri/status/1397288588191178753].

[144] *See, e.g.*, Military-Az Forum, https://www.military-az.com/forum/viewtopic.php?t=2085&start=210, [http://web.archive.org/web/20210926041124/https://www.military-az.com/forum/viewtopic.php?t=2085&start=210].

[145] *See, e.g.*, Defence.Az, *Azerbaijan Army eliminate commander of tank battalion of Armenia*, (Oct. 05, 2020, 14:37), https://bit.ly/3BEaFqw [http://web.archive.org/web/20210926041244/https://defence.az/en/news/146289/azerbaijan-army-eliminated-commander-of-tank-battalion-of-armenia].

[146] *Armenia seeks Russian forces on Azerbaijan border amid tensions*, Al-Jazeera (July 29, 2021), https://www.aljazeera.com/news/2021/7/29/armenia-seeks-russian-forces-on-azerbaijan-border-amid-tensions.

[147] İctimai Television, RADIUS, https://itv.az/tvshows/6.

[148] *Interview (Exclusive)*, ORDU.AZ, (July 9, 2019, 12:58 AM), https://ordu.az/az/news/151791/radius-artiq-unikal-ve-tekrari-olmayan-bir-mehsul-kimi-efirde-oz-yerini-berkidib-musahibe-ekskluziv.

[149] *See* Heydar Mirza, LinkedIn, https://az.linkedin.com/in/heydar-mirza-552a1088.

[150] Caliber.az, Как собственная армия изнасиловала Армению..., YOUTUBE (Mar. 17, 2021), https://www.youtube.com/watch?v=REKYTv9Isss, [http://web.archive.org/web/20210926041901/https://www.youtube.com/watch?v=REKYTv9Isss]; Caliber.az, МО Армении: жив, не жив..., YouTube (Mar. 23, 2021), https://www.youtube.com/watch?v=C-C81iW_ZVrQ, [http://web.archive.org/web/20210926041945/https://www.youtube.com/watch?v=CC81iW_ZVrQ].

[151] VK Terms of Service, *supra* note 43.

[152] *Id.,* at ¶ 6.3.4a

[153] *Id.,* at ¶ 6.3.4d

[154] *Id.,* at ¶ 6.3.4e

[155] *Id.,* at ¶ 6.3.4f

[156] *Id.,* at ¶ 6.3.4k, *see also* section 6. Obligations of the User.

[157] VK Safety Guidelines, Platform Standards, *supra* note 40.

[158] *Id.*

[159] *Id.*.

[160] *Id.*

[161] Evidence of malicious intent includes: (1) animosity based on certain characteristics or differences; (2) offensive behavior, contempt toward other people's values or views; or (3) expression of personal superiority, accompanied by a baseless and unfair attitude toward a specific individual or group of people. *See* VK Safety Guidelines, *supra* note 40.

[162] *Id.*

[163] *Id.*

[164] *Id.*

[165] UN Guiding Principles, *supra* note 3.

[166] GC 34, *supra* note 9.

[167] UN Hate Speech Strategy, *supra* note 34.

[168] ICCPR, *supra* note 7.

[169] VK Safety Guidelines, *supra* note 41

[170] *Id.*

[171] ICCPR, *supra* note 7.

[172] Rabat Plan, *supra* note 35, at ¶ 29.

[173] *Id.* at ¶ 29

[174] *See, e.g.,* Case decision 2021-001-FB-FBR (Facebook Oversight Board 2021), https://www.oversightboard.com/sr/decision/2021/001/pdf-english.

[175] Aisha Jabbarova, *President Aliyev inaugurates Military Trophy Park in Baku,* Azeri Times (2021) https://azeritimes.com/2021/04/12/president-aliyev-inaugurates-military-trophy-park-in-baku/.

[176] In public speeches, Aliyev has described Armenians as vermin; *See* Joe Nerssessian, *The Mixed Messaging of Ilham Aliyev*, EVN Report (Oct. 22, 2020) https://www.evnreport.com/politics/the-mixed-messaging-of-ilham-aliyev.

[177] *See e.g.*, Caliber.az, YouTube, IDEF'21: ВПК Турции продолжает удивлять весь мир… (August 26, 2021) https://www.youtube.com/watch?v=jq5U3Ws82PE; Polygon Azerbaijan, VK (Aug. 27, 2020 7:28 PM), https://vk.com/wall-66054882_51646 [http://web.archive.org/web/20210929005204/https://vk.com/wall-66054882_51646].

[178] *See e.g.*, Military-Az Forum, https://www.military-az.com/forum/viewtopic.php?t=2085&start=210.

[179] United Nations Human Rights Office of the High Commissioner, One-pager on "incitement to hatred", https://www.ohchr.org/sites/default/files/Rabat_threshold_test.pdf.

[180] VK is the most popular social media network in Azerbaijan and has 100 million global monthly users; Vincenzo Cosenza, *The map of social networks in the world - January 2021*, Vincos Blog (Jan. 19, 2021), https://vincos.it/2021/01/19/la-mappa-dei-social-network-nel-mondo-gennaio-2021/.

[181] *See e.g.,* Defence.Az, *Azerbaijan Army eliminate commander of tank battalion of Armenia*, (Oct. 05, 2020, 14:37), https://bit.ly/3BEaFqw [http://web.archive.org/web/20210926041244/https://defence.az/en/news/146289/azerbaijan-army-eliminated-commander-of-tank-battalion-of-armenia].

[182] *See e.g.,* A•shot News 2, VK (March 23, 2021), https://vk.com/wall-169586838_6126 [http://web.archive.org/web/20210929010227/https://vk.com/wall-169586838_6126].

[183] *See e.g.,* Twitter, https://twitter.com/ScourgeOfTengri/status/1397288588191178753, http://web.archive.org/web/20210925214033/https://twitter.com/ScourgeOfTengri/status/1397288588191178753.

[184] *See e.g.,* Military-Az Forum, https://www.military-az.com/forum/viewtopic.php?t=2085&start=210, [http://web.archive.org/web/20210926041124/https://www.military-az.com/forum/viewtopic.php?t=2085&start=2].

[185] Their most recent video was posted Sept. 24, 2021; Caliber.az, YouTube, Перевооружение ВМС Турции и геополитика Средиземноморья, Часть 2, (Sept. 24, 2021) https://www.youtube.com/watch?v=ir7-_wK2RHI.

[186] Narratives employed in the Rwandan genocide include employing dehumanizing language, fear-mongering, messages of inherent inferiority or superiority, encourage a lack of empathy, and expressions of "visceral scorn"; *See* The Prosecutor v. Ferdinand Nahimana, Jean Bosco Baryagwiza and Hasan Ngeze, Case No. ICTR-99-52-T (Int'l. Crim. Trib. for Rwanda December 3, 2003).

[187] *See, e.g.*, Hikmet Hajiyev, Assistant of the President of the Republic of Azerbaijan, Head of Foreign Policy Affairs Department of the Presidential Administration (@HikmetHajiyev), Twitter (Oct. 11, 2020, 1:24 AM), https://twitter.com/HikmetHajiyev/status/1315161389372243968 [https://web.archive.org/web/20201011062912/https://twitter.com/HikmetHajiyev/status/1315161389372243968]; Hikmet Hajiyev (@HikmetHajiyev), Twitter (July 4, 2020, 10:00 AM), https://twitter.com/HikmetHajiyev/status/1279414780382633986 [https://web.archive.org/web/20200704141331/https://twitter.com/HikmetHajiyev/status/1279414780382633986].

[188] Armenian Occupation Watch (@ArmenOccupWatch), Twitter (Oct. 7, 2020, 12:23 PM), https://twitter.com/ArmenOccupWatch/status/1313877651682189314 [https://web.archive.org/web/20201007170332/https://twitter.com/ArmenOccupWatch/status/1313877651682189314].

[189] Twitter (Oct. 9, 2020), https://twitter.com/shahlam_/status/1314461514913456130 [https://web.archive.org/web/20201009070403/https://twitter.com/shahlam_/status/1314461514913456130].

[190] *Synthetic and Manipulated Media Policy*, Twitter, Wayback Machine (Sept. 27, 2020), https://web.archive.org/web/20200927194248/https://help.twitter.com/en/rules-and-policies/manipulated-media [hereinafter Twitter Synthetic and Manipulated Media Policy Sept. 2020]; *Synthetic and Manipulated Media Policy,* Twitter, Wayback Machine (Nov. 13, 2020), https://web.archive.org/web/20201113023157/https://help.twitter.com/en/rules-and-policies/manipulated-media [hereinafter Twitter Synthetic and Manipulated Media Policy Nov. 2020].

[191] *Id.*

[192] *Id.*

[193] *Id.*

[194] *Id.*

[195] *Id.*

[196] *Id.*

[197] *Id.*

[198] *Id.*

[199] *Id.*

[200] *Id.*

[201] Twitter Synthetic and Manipulated Media Policy, *supra* note 190.

[202] Khan 2021 Report, *supra* note 8, at ¶ 42.

[203] Twitter Synthetic and Manipulated Media Policy Sept. 2020, *supra* note 190; Twitter Synthetic and Manipulated Media Policy Nov. 2020, *supra* note 190.

[204] Facebook, Armenia-Artsakh Awareness Center, https://www.facebook.com/armenianawareness/posts/133866341822107.

[205] ArmeniaFund, *About Armenia Fund*, https://www.armeniafund.org/about/.

[206] *Armenia/Azerbaijan: Decapitation and war crimes in gruesome videos must be urgently investigated*, Amnesty International (Dec. 10, 2020), https://www.amnesty.org/en/latest/news/2020/12/armenia-azerbaijan-decapitation-and-war-crimes-in-gruesome-videos-must-be-urgently-investigated/ [hereinafter Amnesty Decapitation Report].

[207] *Id.*

[208] Twitter (Nov. 4, 2020, 5:37 PM), https://twitter.com/Jake_Hanrahan/status/1324118526551117826 [https://web.archive.org/web/20201104223830/https://twitter.com/Jake_Hanrahan/status/1324118526551117826].

[209] Amnesty Decapitation Report, *supra* note 206.

[210] *Violent and Graphic Content*, Transparency Center (Aug. 27, 2020), https://transparency.fb.com/policies/community-standards/violent-graphic-content/ [hereinafter Facebook Violent and Graphic Policy Aug. 2020]; *Violent and Graphic Content*, Transparency Center (Nov. 18, 2020), https://transparency.fb.com/policies/community-standards/violent-graphic-content/ [hereinafter Facebook Violent and Graphic Policy Nov. 2020].

[211] *Id.*

[212] *Id.*

[213] *Id.*

[214] Facebook Violent and Graphic Content Policy Sept. 2020, *supra* note 210; Facebook Violent and Graphic Content POlicy Nov. 2020, *supra* note 210.

[215] *See An Update to How We Address Movements and Organizations Tied to Violence*, About Facebook (Aug. 19, 2020), https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/.

[216] *See Violence and Incitement*, Transparency Center (Sept. 3, 2020), https://transparency.fb.com/policies/community-standards/violence-incitement/ [hereinafter Facebook Violence and Incitement Policy Sept. 2020]; *Violence and Incitement*, Transparency Center (Nov. 18, 2020), https://transparency.fb.com/policies/community-standards/violence-incitement/ [hereinafter Facebook Violence and Incitement Policy Nov. 2020].

[217] *Cf.* Case Decision 2020-003-FB-UA (Facebook Oversight Board 2021), https://oversightboard.com/decision/FB-QBJDASCV/.

[218] Facebook Violence and Incitement Policy, Sept 2020, *supra* note 216; Facebook Violence and Incitement Policy Nov. 2020, *supra* note 216.

[219] *See, e.g.*, Oversight Board Brazil COVID-19 decision, *supra* note 119; Case Decision 2020-006-FB-FBR (Facebook Oversight Board 2021), https://oversightboard.com/decision/FB-XWJQBU9A/.

[220] *Id.*

[221] *See e.g.* Astroturfing Report, *supra* note 120.

[222] Instagram, Karabakh is Azerbaijan, https://www.instagram.com/karabakhisazerbaijann_/.

[223] Facebook, Karabakh is Azerbaijan, https://www.facebook.com/karabakhisazerbaijanmedia/.

[224] Twitter, Karabakh is Azerbaijan, https://twitter.com/karabakh_isaze?lang=en.

[225] YouTube, Karabakh is Azerbaijan, https://www.youtube.com/c/KarabakhisAzerbaijanMedia/about.

[226] Telegram, Karabakh is Azerbaijan, https://t.me/KarabakhisAZE.

[227] YouTube, Karabakh is Azerbaijan, https://www.youtube.com/c/KarabakhisAzerbaijanMedia/about.

[228] Baku Engineering University, https://beu.edu.az/en/page/bmu-haqqinda-64.

[229] Instagram (Oct. 3, 2020, 11:35 AM), https://www.instagram.com/p/CF4vdvlA4Ja/.

[230] Hraparak TV, *Azeris spread false information using the name "Hraparak,"* Hraparak.am (Sept. 27, 2020), https://hraparak.am/post/24a00c25aeae34e397654e676d76a3b2. (Warning that the account operated by "hraparak_tv" is fake, designed to look like the official Instagram account that belongs to the actual Hraparak TV media outlet). This article is originally in Armenian but was translated using Google Translate.

[231] YouTube (Oct. 2, 2020, 3:13 AM), https://www.youtube.com/watch?app=desktop&v=Lvoi-KZuDgIA.

[232] Twitter (Oct. 1, 2020, 1:56 PM), https://twitter.com/africaken1/status/1311726589542117376 [https://web.archive.org/web/20201003000652/https://twitter.com/africaken1/status/1311726589542117376].

[233] *See* YouTube (Oct. 1, 2020, 6:51 AM), https://www.youtube.com/watch?v=AK-lOqmIoCk [https://web.archive.org/web/20201107155534/https://www.youtube.com/watch?v=AK-lOqmIoCk&t=2s].

[234] Instagram (Oct. 3, 2020, 11:35 AM), https://www.instagram.com/p/CF4vdvlA4Ja/.

[235] *See*, *e.g.*, *Another trick of Armenians*, Milli.Az (Oct. 2, 2020), https://news.milli.az/politics/886026.html (translated using Google Translate); *Armenians also involved civilians in fighting*, 525.az (Oct. 2, 2020), https://525.az/?name=xeber&news_id=150938 (translated using Google Translate); Trend, *Armenia involving civilians in combat operations against Azerbaijan*, AzerNews (Oct. 2, 2020), https://www.azernews.az/aggression/169840.html (translated using Google Translate).

[236] *See* Facebook Violence and Incitement Policy Sept. 2020, *supra* note 216; Facebook Violence and Incitement Policy Nov. 2020, *supra* note 216.

[237] *Id*.

[238] *See An Update to How We Address Movements and Organizations Tied to Violence*, About Facebook (Aug. 19, 2020), https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/ (last updated Jan. 19, 2021) [hereinafter FB policy update].

[239] *Hard Questions: What's Facebook's Strategy for Stopping False News?,* Wayback Machine (Aug. 16, 2020), https://web.archive.org/web/20200816052419/https://about.fb.com//news/2018/05/hard-questions-false-news/.

[240] *See*, *e.g.*, Oversight Board Brazil COVID-19 decision, *supra* note 119.

[241] *Cf.* Case Decision 2020-003-FB-UA (Facebook Oversight Board 2021), https://oversightboard.com/decision/FB-QBJDASCV/.

[242] Khan 2021 Report, *supra* note 8, at ¶ 42.

[243] Facebook Violence and Incitement Policy Sept. 2020, *supra* note 216; Facebook Violence and Incitement Policy Nov. 2020, *supra* note 216.

[244] Facebook, *False News*, https://transparency.fb.com/policies/community-standards/false-news/.

[245] Instagram does not provide information on whether a user's account was suspended in the past. However, it is clear that KIAz's account was not suspended during the conflict because content was posted to their account several times a day on almost every single day of the conflict, which would not have been possible if their account had been suspended for a short duration. Also, an archive on the Wayback Machine from November 6, 2020 reveals that the account was still fully operational at that time. *See* Wayback Machine (Nov. 6, 2020), https://web.archive.org/web/20201106072556/https://www.instagram.com/karabakhisazerbaijann_/.

[246] This aligns with the "responsibility to respect" approach advocated by former Former UN Special Rapporteur on freedom of opinion and expression, David Kaye, *See e.g.* Kaye 2018 Report, *supra* note 3, ¶ 1.

[247] GC 34, *supra* note 9, ¶ 25

[248] *Id.*

[249] In particular, Facebook and Instagrams policies are scattered throughout their Terms of Service and elsewhere, including blog posts. The Facebook Oversight Board has held that this should be addressed. *See e.g.* Case Decision 2020-006-FB-FBR (Facebook Oversight Board 2021), https://www.oversightboard.com/decision/FB-XWJQBU9A. If not listed under their Community Standards or Community Guidelines respectively, then Facebook and Instagram's policies can be found on Facebook's Newsroom.

[250] GC 34, *supra* note 9, at ¶¶ 33-34.

[251] Working Group, *supra* note 4.

[252] *Id.* at ¶ 46

[253] *Id,* at ¶ 47

[254] *Id.* at ¶ 48

[255] *Id.* at ¶ 65

[256] *Id.* at ¶ 64

[257] *Id.* at ¶ 65

FOLLOW US

@promiseinstUCLA

WRITE US

promiseinstitute@law.ucla.edu

385 Charles E. Young Drive, East
Los Angeles, California 90095