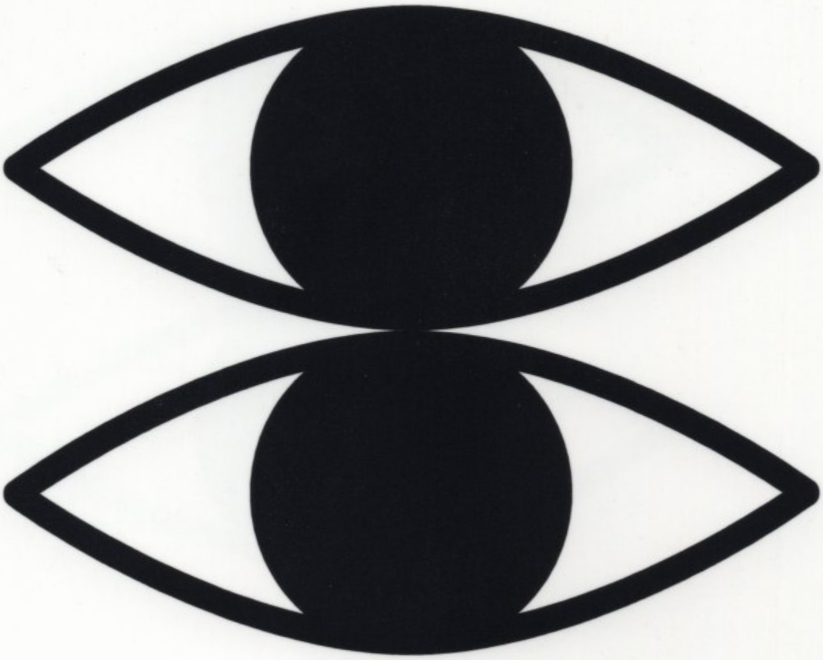
Reading Like a Computer



A Semantic Guide to Facebook's Content Moderation Policies

unknown
UNKNOWNS

foreward

In 2017, the Guardian published leaked Powerpoint slides for training Facebook content moderators.[1] The contents of this book are all taken from these training documents. None of the data has been embellished or editorialized. I have used an asterix and italics font to denote notations where the logical argument was unresolved or uncertain.

Terms like "protected" and "vulnerable" groups are present in training materials but their definitions are ahistorical and unrelated to legal precedent.[2] In Facebook's alternate universe, white people and black people are equally in need of protection.[3] Christians, Muslims, Wiccans, and Satanists are interchangeable as protected religious groups. While one may argue that this equation protects all people from being targets of hate, that is not how it works out. Occupations and social classes are not protected. Their non-protected status can outweigh the protection of another group they are part of. This means Muslims are protected from hate but Muslim cab drivers are not.[4] Criminals (both alleged and convicted) can have horrific violent fantasies projected on them (Facebook calls this "celebrating 'justice'"). When users describe immigrants as criminals, they can sidestep the protections for this group.

I study computational linguistics, programming classifiers that automate the tagging of online texts. I see no horizon where an automated algorithm will be able to distinguish the kind of hate that Facebook allows ("aspirational," "conditional," "sarcastic") from that which is not allowed (a combination of factors such as target and method or "dehumanizing"). The infinite meanings derived from combinations of text and image compound this challenge. For the foreseeable future, applying these confusing guidelines will be the labor of underpaid and overworked contractors.[5] Facebook's training manuals imply that human contractors can both mentally distance themselves and apply procedural, computer-like thinking to violent and offensive speech.[6]

By combining these disparate Powerpoints out in the wild into a more user-friendly and cohesive document, I am not suggesting that more attention to graphic design would make this job or task easier. Instead, I question Facebook's very premise that some forms of hateful and violent speech contribute to enriching interactions and a stronger community.[7]

-Angie Waller, 2018

[1] Facebook. "The Facebook Files." The Guardian, May 2017. https://www.theguardian.com/news/series/facebook-files

[2] Anna Vlasits. "By Facebook's Logic, Who Is Protected From Hate Speech?" Wired, July 1, 2017. https://www.wired.com/story/facebook-hate-speech-moderation/.

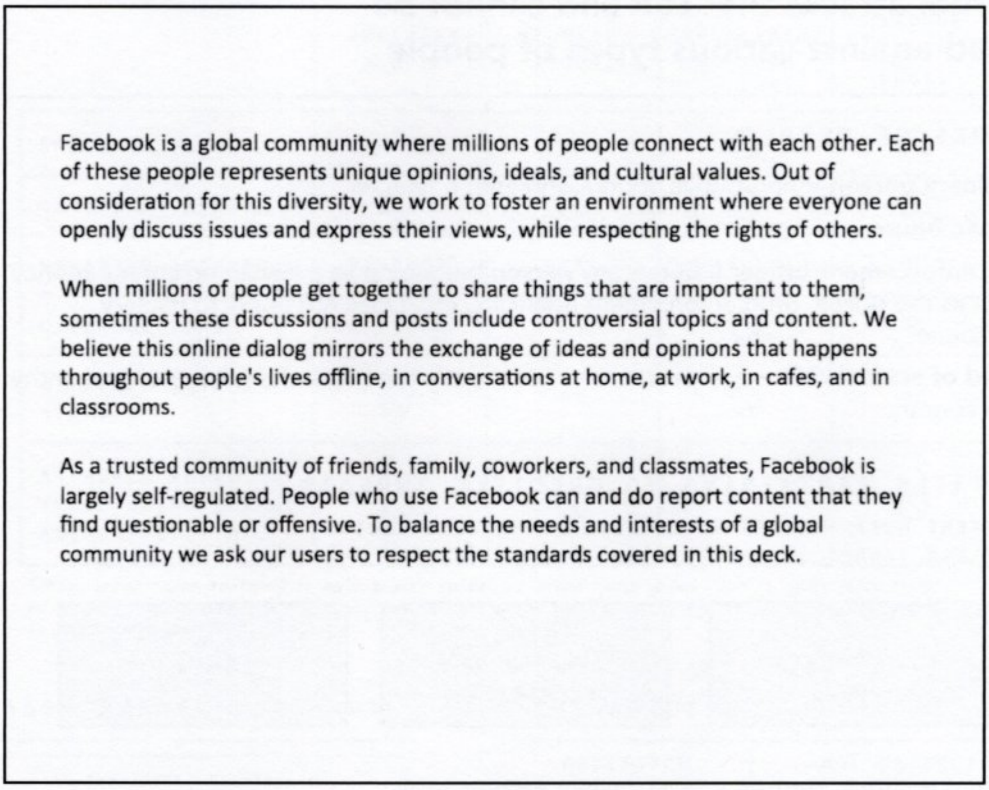
[3] Julia Angwin, Hannes Grassegger. "Facebook's Secret Censorship Rules Protect White Men…." Text/html. ProPublica, June 28, 2017. https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

[4] Simon Adler, Tracie Hunte. "Post No Evil | Radiolab | WNYC Studios." wnycstudios. Accessed September 19, 2018. https://www.wnycstudios.org/story/post-no-evil/.

[5] Olivia Solon. "Underpaid and Overburdened: The Life of a Facebook Moderator." The Guardian, May 25, 2017, sec. News. https://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content.

[6] Anna Vlasits.

[7] Facebook.



Facebook is a global community where millions of people connect with each other. Each of these people represents unique opinions, ideals, and cultural values. Out of consideration for this diversity, we work to foster an environment where everyone can openly discuss issues and express their views, while respecting the rights of others.

When millions of people get together to share things that are important to them, sometimes these discussions and posts include controversial topics and content. We believe this online dialog mirrors the exchange of ideas and opinions that happens throughout people's lives offline, in conversations at home, at work, in cafes, and in classrooms.

As a trusted community of friends, family, coworkers, and classmates, Facebook is largely self-regulated. People who use Facebook can and do report content that they find questionable or offensive. To balance the needs and interests of a global community we ask our users to respect the standards covered in this deck.

SLIDE FROM FACEBOOK CONTENT MODERATION DOCUMENTS
HTTPS://WWW.THEGUARDIAN.COM/NEWS/SERIES/FACEBOOK-FILES

verbal attacks that can and cannot be used against various types of people

TYPES OF PEOPLE

ordinary person = non-public figures who aren't 'famous'

public figure = any person featured in any mass medium (internet, news, etc.)

law enforcement officer (LEO) = any person belonging to a law enforcement agency such as the police, drug enforcement agencies, etc. Does not apply to military personnel

head of state (HOS) = any person who is currently the head of a ruling political entity in a country

DETAILS PERTAINING TO CREDIBLE THREATS

CONTENT MUST SPECIFY 2 OUT OF 3 OF THESE DETAILS
(UNLESS TARGET IS HEAD OF STATE OR POLICE)

TIME/PLACE	METHOD (WEAPON OR DESCRIPTION)	TARGET

time = {tonight, tomorrow, 5pm, "when he gets back",...}

method = {cut throat, stab, shoot,...}

weapon described = {"I have the gun", "with my knife",...}

EXAMPLES

"I'm ready to walk up to Trump Tower and shoot that SOB"

AGENT: I'm — TIME/PLACE: Trump Tower — METHOD: and shoot — TARGET: that SOB

✗ NOT ALLOWED

"I'm going to stab the idiots in France"**

AGENT: I'm — METHOD: stab — TARGET (NOT SPECIFIC): idiots — TIME/PLACE: France

✓ ALLOWED

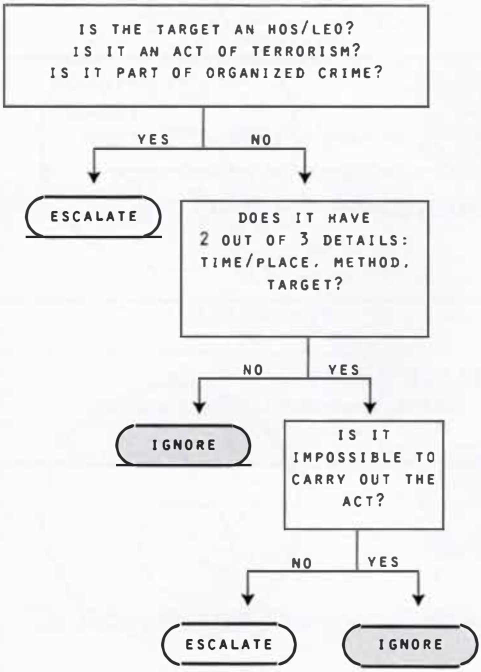
It is not clear in the context of this example if "that SOB" is assumed to be the President of the U.S. based on the Trump Tower location. Regardless, a single person in a specific circumscribed location is the target, so criteria are met.

***In this example, more than one idiot is referred to so the target is non-specific. The geographic boundaries of France fall outside of the parameters for threats on location which are discussed in more detail on p.11.*

	ORDINARY PEOPLE	PUBLIC FIGURE	LAW ENFORCEMENT OFFICER (LEO)	HEAD OF STATE (HOS)
EMPTY THREATS	DEPENDS*	OK	⊘	⊘
CREDIBLE THREATS	⊘	⊘	⊘	⊘
NEGATIVE REFERENCE	DEPENDS*	OK	⊘	OK
CYBERBULLYING	DEPENDS*	OK	⊘	OK
ATTACKED WITH HATE SYMBOLS	⊘	OK	⊘	OK
ATTACKED BASED ON THEIR BEING SEXUAL ASSAULT VICTIM	⊘	⊘	⊘	⊘

**These areas refer trainees to hate speech rules. It seems these areas require extra parameters to be met, mostly that the target of the threat is the one tagged in the post and the one who is reporting the threat.*

ASSESSMENT FLOWCHART


IS THE TARGET AN HOS/LEO?
IS IT AN ACT OF TERRORISM?
IS IT PART OF ORGANIZED CRIME?
YES
NO
ESCALATE
DOES IT HAVE 2 OUT OF 3 DETAILS: TIME/PLACE, METHOD, TARGET?
NO
YES
IGNORE
IS IT IMPOSSIBLE TO CARRY OUT THE ACT?
NO
YES
ESCALATE
IGNORE

SCENARIOS FOR INDIVIDUALS

PROTECTED	NOT PROTECTED
current president	former president
presidential candidate	previous presidential candidate
rape victim	criminal
police	former military personnel

protected groups, quasi-protected groups, and non-protected modifiers

INDIVIDUALS

GROUPS

HUMANS

PROTECTED GROUPS

race = {white, black, hispanic, asian, ...}

ethnicity = {American, Indians, Aborigines, ...}

national origin = {Americans, British, French, Chinese, ...}

religion = {Catholics, Protestants, Muslims, Sunni, Shia, Scientologists, Mormons, Jehovah's witnesses, Satanists, Atheists...}

sex={male, female}

gender identity = {heterosexual, bisexual, homosexual, asexual}

sexual orientation = {lesbian, gay, bisexual, transgender}

disability = {physical, sensory, intellectual, mental, developmental}

serious disease = {any life threatening disease}

"WE PROTECT THE FOLLOWERS OF A RELIGION. NOT THE RELIGION ITSELF."

VULNERABLE PERSON

{heads of state, next in line for head of state, candidates for head of state, specific law enforcement officers, witnesses and informants, people with history of assassination attempts, people listed as targets on Hit Lists created by Banned Dangerous Orgs., activists and journalists}

VULNERABLE GROUPS

GLOBALLY PROTECTED	PROTECTED IN PHILIPPINES
{homeless people, foreigners, Zionists}	{drug dealers, drug users, drug addicts}

NOT PROTECTED CATEGORIES

social class = {rich, poor, middle class, working class,...}

appearance = {blonde, brunette, short, tall, fat, thin,...}

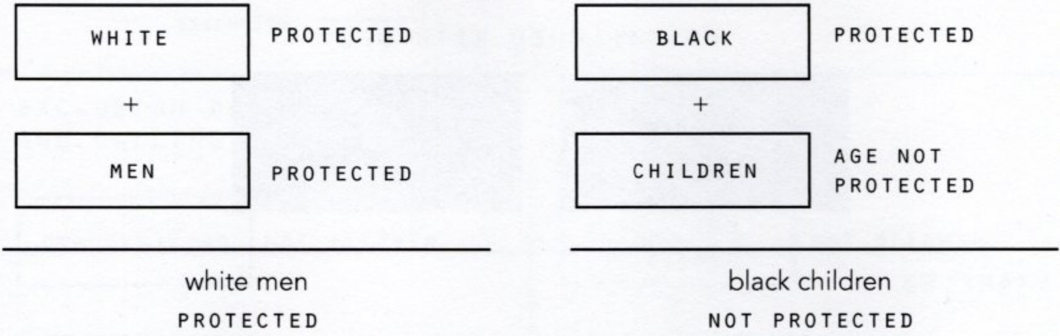
political ideology = {Republicans, Democrats, Socialists, Communists, Revolutionaries,...}

countries = {Ireland, England, France, USA, Brazil, Spain,...}

NON-PROTECTED MODIFIERS*

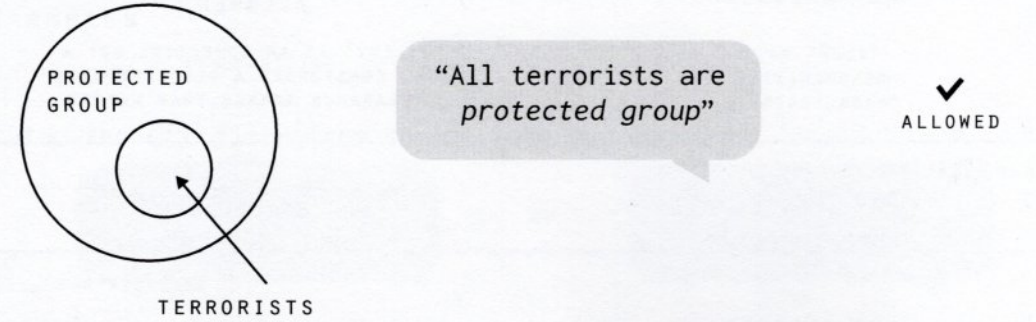
not protected modifiers = {occupation, age, *not protected categories*}

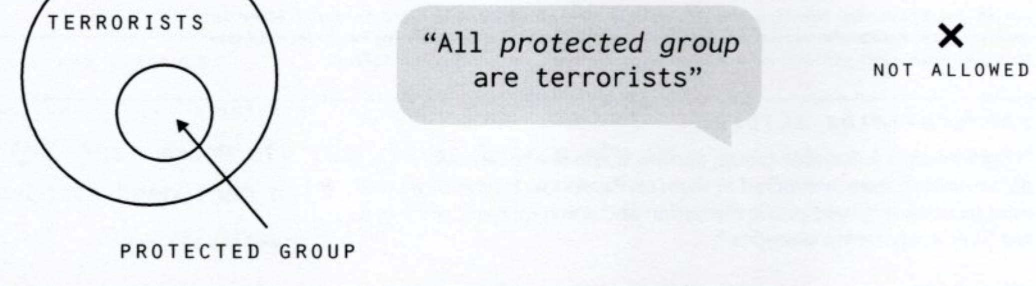
**When a protected category is combined with a non-protected modifier, the non-protected modifier outweighs the protected modifier even if the subset is part of the protected set.*



NON-PROTECTED MODIFIER AND WORD ORDER**

***When making a generalization about a group of people, subsets that are a non-protected category cancel out the protections of the set that contains them.*





quasi-protected and non-protected categories

QUASI-PROTECTED CATEGORY (QPC)

QPC = {migrants, refugees, immigrants, asylum seekers,...}

NOT ALLOWED WITH QPC

CALLS FOR VIOLENCE

ASSIGNING DEHUMANIZING CHARACTERISTICS

EXAMPLES

DEHUMANIZING CHARACTERISTICS		EDGE CASES
"*QPC* are scum"	vs	"*QPC* are so filthy"
NOT ALLOWED		ALLOWED
"SCUM" IS A DEHUMANIZING CHARACTERISTIC		"FILTHY" IS AN ADJECTIVE NOT A NOUN, CONSIDERED A DESCRIPTION OF APPEARANCE RATHER THAN NATURE
"*QPC?* They're all rape-fugees!"	vs	"*QPC* are thieves and robbers"
NOT ALLOWED		ALLOWED*
DEHUMANIZING CHARACTERISTIC		"DISMISSING" AN ENTIRE QPC CAN BE IGNORED

**It is not clear why labeling the QPC with a made up derogatory word is not allowed but calling them criminals matter-of-factly is ok, especially considering that thieves and criminals can have violent acts wished upon them on the Facebook platform.*

FROM TRAINING SLIDES:

"Migrants are a vulnerable group, and we would like to remove dehumanizing speech directed at them on Facebook. However, we also want to allow for broad public discussion about immigration, which is a hot topic in upcoming elections."

EXCLUSION IS ALLOWED

exclusion = {"should not be allowed", "deport", "build a fence", "keep them out", ...}

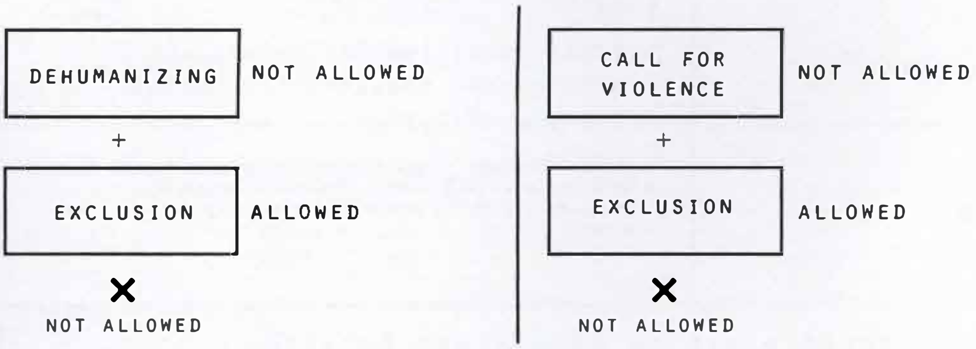
EXAMPLE

"*QPC* should not be allowed into the country."

EXCLUSION

✓ ALLOWED

EXCLUSION DOES NOT OVERRIDE DEHUMANIZING AND CALLING FOR VIOLENCE



DEHUMANIZING NOT ALLOWED

+

EXCLUSION ALLOWED

✗ NOT ALLOWED

CALL FOR VIOLENCE NOT ALLOWED

+

EXCLUSION ALLOWED

✗ NOT ALLOWED

EXAMPLES

DEHUMANIZING

"Stop the *QPC* filth from coming into our country"

EXCLUSION

✗ NOT ALLOWED

CALL FOR VIOLENCE

"Sterilize the *QPC* or else keep them out"

EXCLUSION

✗ NOT ALLOWED

CALL FOR VIOLENCE

DISMISSING

"*QPC* just mooch off the state, that's why we need to keep them out"

EXCLUSION

✓ ALLOWED

crimes and geography specific regulations

TYPES OF CRIME

crimes recognized by facebook

THEFT

ROBBERY

FRAUD

MURDER

VANDALISM

RAPE

regional crimes not recognized by facebook

CLAIMS ABOUT SEXUALITY

PEACEFUL PROTESTS AGAINST GOVERNMENTS*

GENDER SPECIFIC RESTRICTION IN COUNTRIES

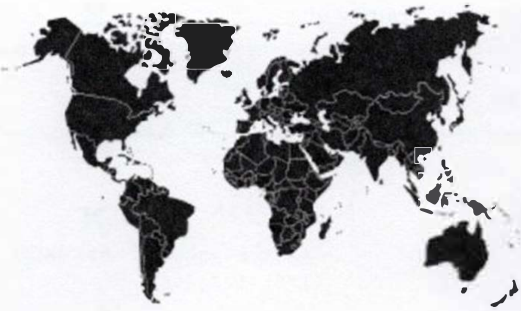
DISCUSSING HISTORICAL EVENTS AND CONTROVERSIAL SUBJECTS (INCLUDING HOLOCAUST DENIAL)

**based on example of arrest around stealing a confederate flag, a person arrested for peaceful protest is recognized as a criminal and can be the subject of "celebrating justice" - see next page*

CONTENT RULES BASED ON GEOGRAPHIC LOCATION

geoblocking = the content is not violating but is hidden from users based on location

international compliance = content is removed based on international regulation



**Examples of how drug dealers are protected was not provided; however, in 2016 Facebook allowed the publication of President Duterte's list of 159 government employees allegedly linked to the country's massive drug trade so that "justice" could be served.*

Source: Erin Hale. "Facebook Fired Up After Philippines' Duterte Releases 'Narco' List." Forbes.

EXAMPLES

1 France, Germany, Israel, Austria

GEOBLOCKING

Holocaust-denial: illegal in 14 countries. "we only consider it for the 4 countries that actively pursue the issue with us." Posts that deny the holocaust are geoblocked in these countries.

2 International

ALL REMOVED

images of burning Turkish flag

maps of Kurdistan

photos and/or text making fun of/ attacking/depicting negatively/ criticizing Atatürk

3 Philippines

REGIONAL RULE

drug dealer. drug user, and "drug addict" are vulnerable groups and lower threshold requirements for credible violence

CREDIBILITY OF THREAT BASED ON SUBJECT AND METHOD

"I'll destroy the Facebook Dublin office."

AGENT

VIOLENT ACT (NOT SPECIFIC)

ORGANIZATION (PLACE SPECIFIC)

✗ NOT ALLOWED

"Destroy the Facebook Dublin office"

VIOLENT ACT (NOT SPECIFIC)

AGENT (NOT SPECIFIC)

ORGANIZATION (PLACE SPECIFIC)

✓ ALLOWED

"Bomb the Facebook Dublin office."

VIOLENT ACT (SPECIFIC)

AGENT (NOT SPECIFIC)

ORGANIZATION (PLACE SPECIFIC)

✗ NOT ALLOWED

CELEBRATING "JUSTICE"

(ADVOCATING VIOLENCE ON A SUSPECTED OR CONVICTED CRIMINAL)

Advocating for capital or cruel and unusual punishment for crimes recognized by Facebook is allowed.*

EXAMPLES

"Pedophiles are going to experience the electric chair!"

CRIMINAL

VIOLENT PUNISHMENT

✓ ALLOWED

"Good to see that bastard hang" context: story about death penalty

CRIMINAL

VIOLENT PUNISHMENT

✓ ALLOWED

*types of violence included state death penalty methods and also vigilante justice such as hanging and shooting.

violence that is humorous, sarcastic, or fictitious

HYPERBOLIC VIOLENCE

Threats using figurative or humorous speech are allowed.

EXAMPLES

"That was a good prank. I'm going to kill you guys!" ✓ ALLOWED

CONTEXT IS PRANK = HUMOROUS

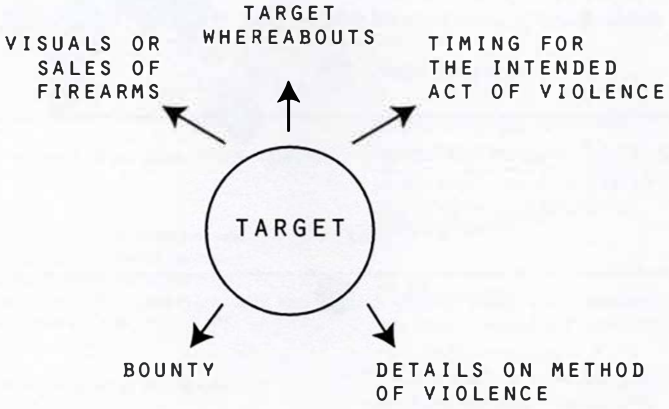
"I am going to disembowel you with a spoon" ✓ ALLOWED

DISEMBOWEL WITH SPOON = FIGURATIVE

THREATS THAT ARE NOT ACHIEVABLE ARE ALLOWED

ACHIEVABLE	NOT ACHIEVABLE
Today	When pigs fly
Tomorrow	When hell freezes over
In 3 hours	Violence in fictional setting
Next time I see you	Violence on someone who is deceased
When it rains	Violence on fictional character or avatar
Sooner or later	

GLOBAL CREDIBILITY STANDARD: PRINCIPLES


VISUALS OR SALES OF FIREARMS
TARGET WHEREABOUTS
TIMING FOR THE INTENDED ACT OF VIOLENCE
TARGET
BOUNTY
DETAILS ON METHOD OF VIOLENCE

violence that is aspirational

Aspirational - Statements that hope for something to happen
{I wish, I hope, I pray, I want}

Conditional - Statements about a possible future action
{if...then..., would..., could...}

ASPIRATIONAL AND CONDITIONAL VIOLENCE AGAINST VULNERABLE GROUPS IS NOT ALLOWED

SEE P.6 FOR EXAMPLES OF VULNERABLE GROUPS

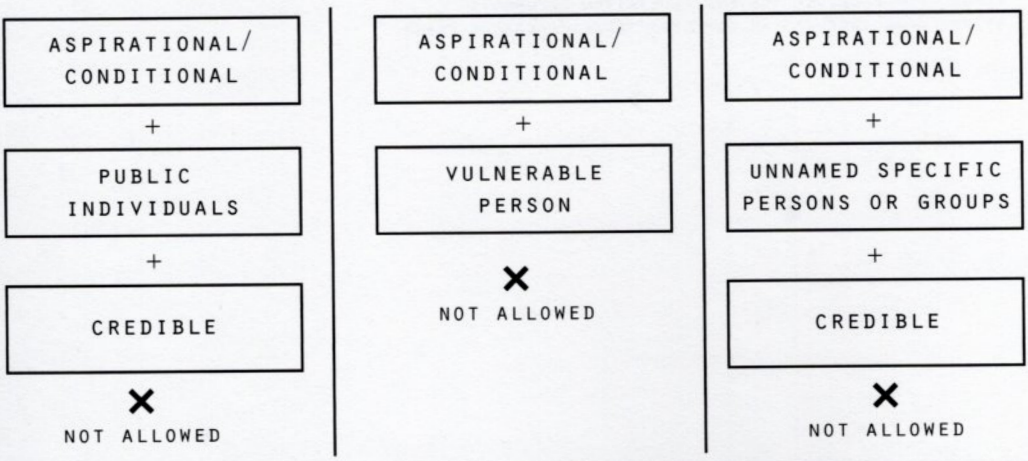
EXAMPLE

"I hate *vulnerable group* and I want to shoot them all"

VIOLENT ACT

✕
NOT ALLOWED

ASPIRATIONAL VIOLENCE AGAINST OTHER CATEGORIES IS NOT ALLOWED WHEN ACCOMPANIED BY A CREDIBLE THREAT



ASPIRATIONAL/ CONDITIONAL + PUBLIC INDIVIDUALS + CREDIBLE ✕ NOT ALLOWED

ASPIRATIONAL/ CONDITIONAL + VULNERABLE PERSON ✕ NOT ALLOWED

ASPIRATIONAL/ CONDITIONAL + UNNAMED SPECIFIC PERSONS OR GROUPS + CREDIBLE ✕ NOT ALLOWED

EXAMPLE

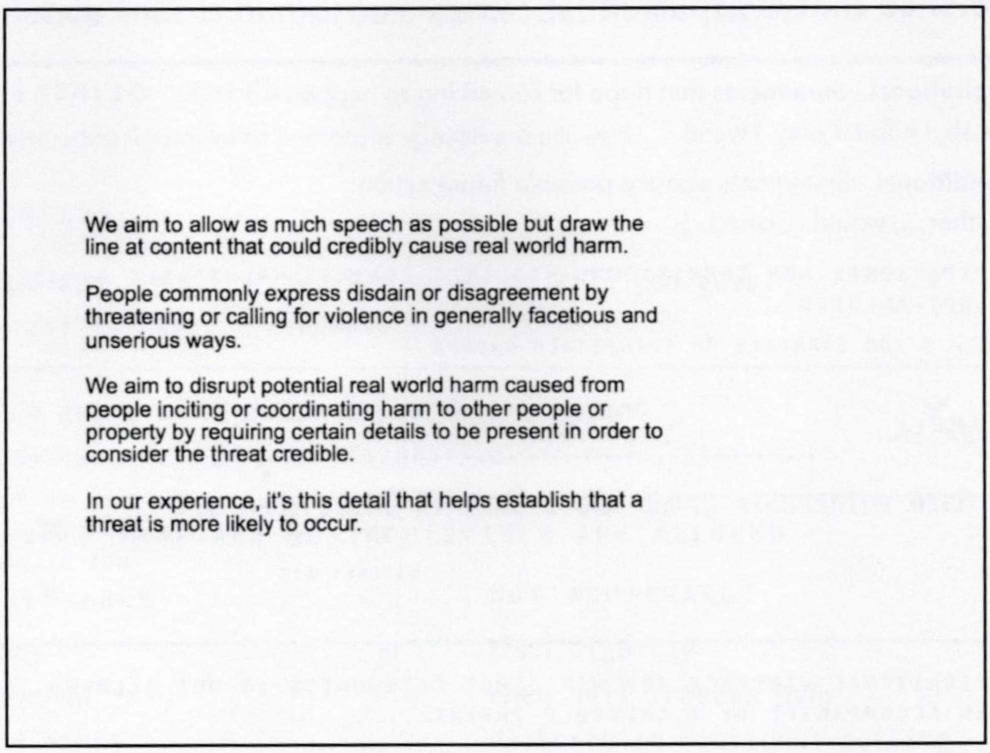
"I hope someone kills you"

ASPIRATION

VIOLENT ACT (NON-SPECIFIC)

UNNAMED SPECIFIC PERSON

✓
ALLOWED



We aim to allow as much speech as possible but draw the line at content that could credibly cause real world harm.

People commonly express disdain or disagreement by threatening or calling for violence in generally facetious and unserious ways.

We aim to disrupt potential real world harm caused from people inciting or coordinating harm to other people or property by requiring certain details to be present in order to consider the threat credible.

In our experience, it's this detail that helps establish that a threat is more likely to occur.

SLIDE FROM FACEBOOK CONTENT MODERATION DOCUMENTS
HTTPS://WWW.THEGUARDIAN.COM/NEWS/SERIES/FACEBOOK-FILES