This document has a list of various themes and thoughts that I had when researching.

Mimicry effect: Boosting positive content rather than deleting harmful content--this is part of what instagram has been doing

Temporal clustering of hate speech: prejudicial crimes are strongly influenced in the short term due to publicized events such as murders committed by a minority, this amplification in hate speech usually lasts about 2 weeks. [1]

- Hashtags can be a good indicator of temporal clustering: ex: #killallmuslims #ferguson #charliehebdo #brussels #banislam #baltimore #mizzou

Backfire effect--Political psychologists Nyhan and Reifler have researched the "backfire effect" where attempts at correcting misperceptions or misinformed beliefs results in firmer beliefs in the misperception or misinformation

- "individuals who receive unwelcome information may not simply resist challenges to their views. Instead they may come to support their original opinion even more strongly" [2]
- It is important to consider the backfire effect in looking at effective forms of intervention/counterspeech, since often presenting facts or engaging in logical/reasoned debate is not the most effective strategy
- A similar psychological phenomenon is "**motivated reasoning**" where people make a strong effort to support the conclusions they seek despite being exposed to contradictory facts

Deindividuation
Depersonalization
Dissociative anonymity
Online disinhibition
Mob-mentality
Flesh Search Engines
justice stewart: "I know it when I see it"

When people perceive cyberbullying or hate speech as accepted by others in their school, workplace, society etc. they are more likely to engage in it themselves.

a significant difference between cyberbullying and hate speech (besides protected groups) is that hate speech can be a one time thing whereas repetition is central to the definition of cyberbullying

what do we do about private hate speech between two racists that aren't attacking anyone directly?

---

[1] King, R.D., and G.M. Sutton. 2013. "High Times for Hate Crime: Explaining the Temporal Clustering of Hate Motivated Offending." Criminology 51 (4): 871–94

[2] Nyhan, Brendan, and Jason Reifler. "When corrections fail: The persistence of political misperceptions." *Political Behavior* 32.2 (2010): 307.

Where is the *hate* in hate speech located? Whose hate is it? Is it the state of mind of the speaker? The state of mind and opinions held by the speaker's audience? Is it the feeling of being hated felt by the victim?

Concerns raised by content moderation:

- Free speech
- Anonymity and privacy
- No platforming
- Further marginalizing groups--example: deleting the account of someone posting anti-racist speech
- Amplifying hate speech by prosecuting/calling it out
- Real name policies have lead to the deactivation of trans folks and drag queens' accounts from facebook, so while attaching a fixed identity may be named as a remedy to online harassment, it can also be a form of discriminatino itself https://cdt.org/blog/facebook-should-reform-its-real-name-policies/