Vaibhav Beohar, W251,
Spring 2020, MIDS

Does online deplatforming of hate speech decrease polarization in societies?

As part of the final project in our field experiments and causal inference course, I propose to study whether *deplatforming* of users generating inflammatory content on social media leads to a general decrease of polarization in societies.

Background and introduction:
Deplatforming is a term used to describe the process of removing someone's access to a channel for delivering (or broadcasting) messages to a wider audience. It may involve not just banning the offending user from the social media site, but also removing their existing and prior content present on the site. These actions restrict the ability of individuals and communities to communicate with each other and with the public. Although deplatforming raises ethical and legal questions, it raises a more pertinent and curious question of whether it is an effective strategy to reduce hate speech and similar calls for violence on social media.

Today, this term has assumed wider connotations and debates because of a hypercharged environment, especially in the wake of the assault on the U.S. Capitol on January 6th 2021 by a large group of protestors who purportedly believed that the November 2020 Presidential elections were rigged. Subsequent to these events, Twitter permanently suspended then U.S. President Donald Trump's personal account. Furthermore, Google, Apple and Amazon suspended Parler, a social media platform favored by the far right.

Deplatorming and similar actions that strip a particular user or group of users of their rights of posting content evoke strong emotions on both sides of the thought spectrum. Some believe that such actions do work by dampening the odds of undesirable social outcomes and unrest (examples are cited from the deplatforming of right-wing provocateur Milo Yiannopoulos who was banned from Twitter in 2016 and Facebook in 2019; and from the removal of Gab, a social media company, from app stores and cloud infrastructure providers). On the other hand, opposing viewpoints believe that deplatforming increases big tech's influence, further sidelines communities and abrogates people's first amendment rights.

Initial approach:
While this topic certainly has a wide scope for research, experiments may be designed on a group of subjects to assess whether disassociation from (or association to) polarizing or inflammatory social media content *(or a group of influencers, or set of inflammatory posts or content),* may lead to an overall reduction (or increase) in biases or prejudices towards a certain set of ideals, principles or thoughts. An initial design would require treatment on a collection of subjects - across a range of ethnicities, age groups, gender identities and social backgrounds - who are at least moderately active on social media platforms, preferably on YouTube or Facebook.

Once such a sample is identified, subjects would need to be classified into two blocks: prejudiced and unprejudiced (or "already polarized" and "non polarized"). This would ensure that chance imbalances of polarized (or non-polarized) subjects, being heavily assigned in either treatment of control groups, are eliminated.

In order to classify our subjects into our two primary blocks (of being prejudiced and unprejudiced), we could start by having folks take a series of implicit bias surveys. One such popular survey would be the Implicit Association Test (IAT) that can assess implicit biases on a variety of topics like politics, religion, race, sexuality, skin-tone, and so forth.

Finally, if sample sizes allow, we could increase the power of our intervention by using a "block within a block" experimental design. For example once subjects are classified into our two primary blocks, further sub-blocks could be constructed on a variety of other prognostic covariates such as gender, age-group, ethnicity, and so forth.

Such a recursive blocking approach would help us design our experiment in a manner that respects the diversity of subjects. We potentially might have subjects scattered around various geographical locations. As such the degree of polarization in coastal areas is probably very different from regions like Idaho or Montana. Therefore, a "cluster within a block" experimental design would not work in our case (because clustering is usually based on regions that share certain similar characteristics, but in reality it can be hard to find cities sharing characteristics).

Therefore the approach here would be to first block on our primary sub-groups (prejudiced and unprejudiced) and then further blocking on such covariates as gender, age-group, income and education levels. This would ensure that a similar number of subjects are assigned to both treatment and control.

Designing the experiment:
The scope of this experiment would focus exclusively on online behavior of subjects. As such, a primary assumption for this study is that responses of subjects reflect their inherent biases and prejudices. Therefore, the experiment involves some key design components: (1) Pre-bias analysis survey, such as IAT described above. Such surveys detect the degree to which our subjects are already "polarized" or not, owing to their prior exposure to online news sources, hate speech, fake news and such. (2) Designing treatment exposure: subjects could be treated with an exposure to alternate forms of online content; to which they have been hitherto unexposed. To this effect, separate online channels could be created on YouTube and/or Facebook to feed our subjects opposite viewpoints for a series of days. It will have to be ensured that our algorithms continuously update with progressively similar videos (hardly a concern given YouTube's much maligned "rabbit hole" algorithms). Furthermore, in order to assess the reaction of participants across a variety of topics, the experiment could also include videos on subjects related to race, economy, LGBTQ-rights and such. (3) Post-bias analysis survey to assess the treatment effect on subjects.

Additional important criteria of randomization, excludability and non-interference have to be carefully considered. As mentioned earlier, after blocking and clustering our subjects, we can randomly assign geographically co-located clusters of subjects within blocks. Concerns about internal validity of our survey owing to exclusion criteria have to be considered by not only urging our subjects to adhere strongly to our survey terms (i.e. restricting social media use and primarily only watching our news channels), but also by standardizing our treatments to make the excludability assumption more likely (by making sure treatment and control groups are

equally assigned and are heterogeneous enough). Finally, because our experiment is conducted via online channels to individuals unknown to each other, we hardly have to be concerned that the non-interference criteria would be breached in our experiment.

Assessment of outcomes:
A variety of metrics are candidates for measures of potential outcomes: reduction in hate speech used in online and social media forums, general acceptance of opposing viewpoints, pre- and post-evaluation of a group of users assigned to treatment and control and evaluation of their inherent biases before and after subjecting them to different viewpoints on social media for a specific duration, and so forth.

However, to make our experimentation simple, the best measure of potential outcome could be very similar to the Clingingsmith, Khwaja and Kremer study, where the authors estimate the impact of winning the lottery to the Hajj pilgrimage on attitudes toward people from other countries.

For example, we first assign a rank to each topic, on a five-point scale of -2 to +2 (-2 being highly polarized and +2 being least). Responses could then be added across topics assessed. We could then assign each subject numeric values (e.g. ranging from -12 being highly polarized, to a +12 being on the opposing swing of the ideological spectrum). We could then treat subjects over a period of time. Subsequent to which, we would assess our subjects again and rank them on the same ordinal scale.

Thus an ATE could be simply calculated via the distribution of responses in the treatment and control groups. As stated earlier, by ensuring a blocked and clustered assignment based on covariates described in earlier sections, we could minimize the noise and variance in the estimated ATE.

Finally, as a parallel observation, a paired samples t-test could also be used to construct a null hypothesis to compare means between our two related groups of samples (prejudiced and unprejudiced). In our field experiment, we would have two values (i.e. pair of values) for the sample (before and after the intervention). In this manner we could evaluate p-values from the paired samples t-test. This would not only ensure an independent and parallel mechanism of evaluation, it would also help us validate the power analysis generated from the original randomization inference experiment.

Additional considerations:
It's worth noting that such online experiments come with their own limitations and risks. For example, we don't know if our sample truly reflects how societies and individuals work. Our subjects could be heavily influenced, not only by social media, but also by family members and friends and other surrounding factors.

Another potential risk could be non-adherence and spill-over. Though our experiment would begin with a strong request to all subjects to stay focused with the treatment conditions, it's highly likely that our subjects would consume regular social media content through their

smartphones, thereby reducing our interventional effects. We would have to devise some ways of preventing such spill-overs. We would need to make sure the control group (unexposed) does not talk to others or they do not find other channels to access the information because we cannot lock them up (which may be a potential risk in many such online experiments).

We also run the risk of not really knowing an ideal duration of conducting such an intervention. How long before we see benefits of deplatforming and non-exposure to polarizing contents in individuals? Does this duration vary between subgroups and individuals? Do individual revealed preferences play any role? If so, how much do they negate any outside effects of social media? Such points (and many more) may impact our study to a certain extent.

Last but not the least, designing isolated social media channels for the treatment and control groups seem to be somewhat challenging and may require time commitment from the group conducting this experiment. It may not be a challenge in settings with non-constrained time limits, but could prove somewhat daunting given our aggressive course schedules.

Conclusion:
Through this study, an attempt (albeit small but an important one) could be made to understand an important question of our times: how much role does social media really play in polarizing individuals and societies? Such questions are bound to be raised in debates focused around free-speech and government regulations. Therefore, our experiment holds great potential to objectively steer such debates with empirical data and analyses.