

Facebook App Research

Connection Integrity Research

# How much of News Feed is Good (or Bad) for the world?

October 23, 2020

REDACTED FOR CONGRESS



## TLDR

Connection integrity has conducted the first international **Good For The World survey**. This survey aims to capture a wide-range of users' subjectively-defined bad experiences on Facebook, supporting both the development of survey-based demotion models and evaluation of integrity efforts.

### Our Findings

9.5% of VPVs are on posts that users feel are "bad for the world" (BFTW) and 63.6% of VPVs are on posts that users feel are "good for the world" (GFTW)

High-reach posts and reshares are more likely to be seen as "bad for the world" than lower-reach and original posts

Civic content is more likely to be rated BFTW than non-civic content

Users in Africa, the Middle East, and Portugal are much more likely to say that posts in their News Feeds are bad for the world than are users in the U.S.

### Our Recommendations

These findings set a baseline for tracking progress toward personalized integrity goals, but careful consideration is needed to establish meaningful goals around these measures.

We should continue working to correct the distribution of content on platform to discourage sharing of harmful content and encourage distribution of good experiences

Civic and personalized integrity teams should partner to reduce bad experiences with civic content while establishing guardrails to protect distribution of essential content

More global efforts on integrity are needed to more equitably serve users across diverse markets

REDACTED FOR CONGRESS

October 23, 2020 · 🌐

1

4 Comments

Like

Comment

Share

Do we have a guide line on what is considered essential content?

Like · Reply · 29w

1

are these statistics weighted?

Like · Reply · 28w

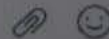
the  
topline statistics are weighted accounting for user country, age, and gender (s5)

Like · Reply · 28w

Thanks. I suspect you'll want some sort of engagement or consumption correction.  
<https://fb.workplace.com/notes/1019048995211342>  
trips-  
and-many-other-surveys-  
under-count-new-and-low-  
vpv-users-leading-to-  
undere/1019048995211342

Like · Reply · 28w

Write a reply...



Write a comment...



Save

## Is this kind of post good for the world?

This is the "Good For The World" (GFTW) survey question.

- Respondents answer on a scale from:
  - Very good
  - Somewhat good
  - Neutral
  - Somewhat bad
  - Very bad
- International survey
  - Translated for 54 locales

Sagi, 2019



Is this kind of post good for the world?

- Very good
- Somewhat good
- Neutral
- Somewhat bad
- Very bad
- There is no post displayed

Continue

REDACTED FOR CONGRESS

October 23, 2020

10 Comments

Like Comment Share

what would you say about this post?

Like · Reply · 29w

3

"Very extremely extraordinarily good for the world"

Like · Reply · 29w

7

I should've warned you you were featured but glad you rate this favorably

Like · Reply · 29w

1

Write a reply...

how do respondents interpret this question? How should they interpret this question?

Like · Reply · 28w

1

As referenced later in the deck, [redacted]'s work examined how respondents think about this question: <https://fb.workplace.com/photo?fbid=490257215241967&set=a.490255551908800> the data on slide 12 also speak to

Write a comment...

interpret this question?

Like · Reply · 28w



[redacted] As  
referenced later in the deck,  
[redacted]'s work  
examined how respondents  
think about this question:  
[https://fb.workplace.com/photo?  
fbid=490257215241967&se  
t=a.490255551908800](https://fb.workplace.com/photo?fbid=490257215241967&set=a.490255551908800) the  
data on slide 12 also speak to  
how responses to this  
question track known  
integrity problems

Like · Reply · 28w

[redacted] As  
for how they "should"  
interpret the question, it is  
meant to be a subjective  
measure of their own  
personal opinions and  
perceptions. More generally,  
the question is meant to  
support demotion models  
that would reduce unwanted  
negative experiences. So one  
hope is that they would  
answer the question about  
the post itself and not the  
content being discussed. E.g.  
you could imagine a post  
about a tragic event isn't  
\_itself\_ BFTW, even if the  
tragedy was. Unless of  
course they thought the post  
itself was bad (e.g. b/c it was  
sensationalizing tragedy, or  
just b/c of too much tragic  
news)

Like · Reply · 28w

Write a comment...



itself was bad (e.g. b/c it was sensationalizing tragedy, or just b/c of too much tragic news)

Like · Reply · 28w

I don't think the deck fully answers the question [redacted] was asking here. I am also curious about the sample characteristics in the qualitative / eye-tracking study as I have a hard time imagining these are international insights.

On how they should interpret the question, this is to capture their own personal opinions and perceptions on what exactly?

Like · Reply · 28w

[redacted] should tag [redacted] and [redacted] who were more involved with the creation of the question and might have a clearer answer for you. But as I mentioned above the context behind the creation of this question was to generate training data for personalized demotion models that would reduce unwanted negative experiences in Feed. The way they operationalized this was in a question asking if the post was good for the world, the reasoning being that people generally wouldn't want to see things they felt were bad for the world. In this way the spirit of the questions is much like the

Write a comment...



should tag [redacted] and [redacted] who were more involved with the creation of the question and might have a clearer answer for you. But as I mentioned above the context behind the creation of this question was to generate training data for personalized demotion models that would reduce unwanted negative experiences in Feed. The way they operationalized this was in a question asking if the post was good for the world, the reasoning being that people generally wouldn't want to see things they felt were bad for the world. In this way the spirit of the questions is much like the Worth Your Time question that drives part of the feed value model on the relevance side of things: the goal is to measure just exactly what the question asks. The way we have tried to validate this beyond simple face validity is through the qual work [redacted] did, as well as examining the correlation between the survey responses and known integrity problem areas. This is challenging though as the question is meant to be subjective, so it is not clear just how strongly correlated with objective problems it ought to be.

Like · Reply · 28w



were bad for the world. In this way the spirit of the questions is much like the Worth Your Time question that drives part of the feed value model on the relevance side of things: the goal is to measure just exactly what the question asks. The way we have tried to validate this beyond simple face validity is through the qual work [REDACTED] did, as well as examining the correlation between the survey responses and known integrity problem areas. This is challenging though as the question is meant to be subjective, so it is not clear just how strongly correlated with objective problems it ought to be.

Like · Reply · 28w

[REDACTED] the initial qualitative work (eye tracking study) was done in English but we did some follow-up qualitative research in additional markets and languages. If you'd like to walk through our work developing and assessing the question itself I'm happy to set up time to talk through that with you.

Like · Reply · 28w



[REDACTED] happy to read the research reports!

Like · Reply · 28w



Save

## GFTW survey represents a "broad brush" approach to integrity

Rather than focusing on prescriptively-defined integrity problem areas, the GFTW survey aims to paint with a broad brush to address a wide range of users' subjectively defined bad experiences

REDACTED FOR CONGRESS



Save

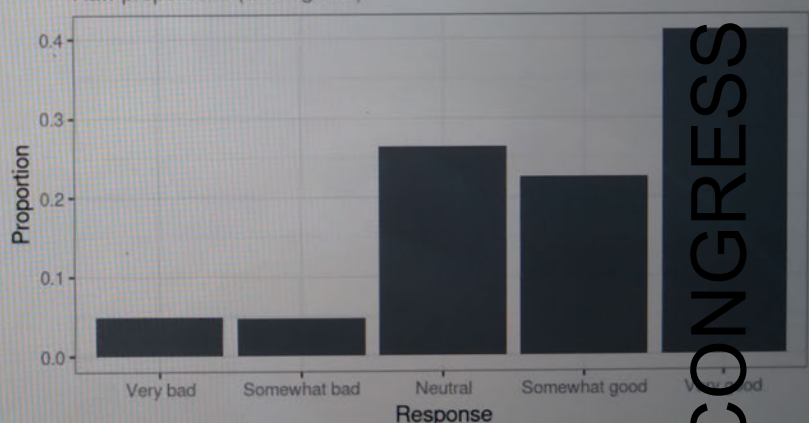


## 9.5% of VPVs are on posts that users feel are "bad for the world"

Respondents rated 9.5% of the posts in their News Feeds as "very" or "somewhat" bad for the world (CI<sub>95</sub> [9.4% to 10%])

- In comparison, respondents rated 63.6% of posts in their News Feeds as "somewhat" or "very" good for the world
- Estimates are weighted to Facebook DAP adjusting for users' gender, age, and country
- International survey
  - n = 143,511 users
  - N = 517,625 GFTW responses
- Survey fielded from Aug. 14 to Sept. 10, 2020

GFTW survey responses  
Raw proportions (unweighted)



REDACTED FOR CONGRESS

October 23, 2020 · 🐾

8 Comments

👍 Like    💬 Comment    ➦ Share

██████████ This is a random sample of posts?

Like · Reply · 29w

██████████ Random sample of VPVs

Like · Reply · 29w

👍 1

██████████ Is the lower bound on the 95% CI around 9.5 really 9.4?

Like · Reply · 28w

n = 517k

ratings

Like · Reply · 28w

██████████ But the upper bound is 10.

Like · Reply · 28w

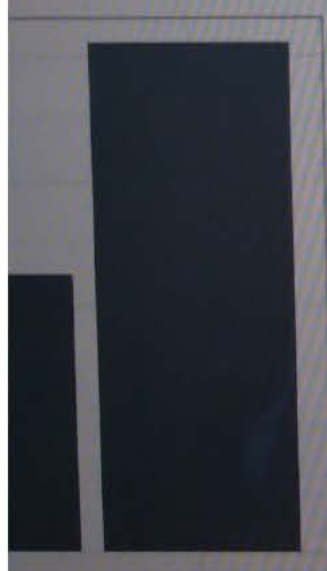
██████████ I'm just curious why you'd have an asymmetric CI in this case.

Like · Reply · 28w

██████████ It's a binomial proportion so it wasn't surprising to me it could be asymmetrical given it's near the bound. But now I'm realizing I'm not 100% sure what's going on under the hood: I'm computing these using the survey

rd"

ld



good    Very good

Random sample of VPVs

Like · Reply · 29w



Is the lower bound on the 95% CI around 9.5 really 9.4?

Like · Reply · 28w

n = 517k

ratings

Like · Reply · 28w

But the upper bound is 10.

Like · Reply · 28w

I'm just curious why you'd have an asymmetric CI in this case.

Like · Reply · 28w

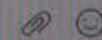
It's a

binomial proportion so it wasn't surprising to me it could be asymmetrical given it's near the bound. But now I'm realizing I'm not 100% sure what's going on under the hood: I'm computing these using the survey package which may be doing some kind of bootstrapping for these binomial proportions. I can dig around a bit in the docs

Like · Reply · 28w



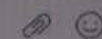
Write a reply...



I would also consider adjusting for exposure to civic content in your bias correction

Like · Reply · 28w

Write a comment...



02  
—

# What kinds of posts are Bad For The World?

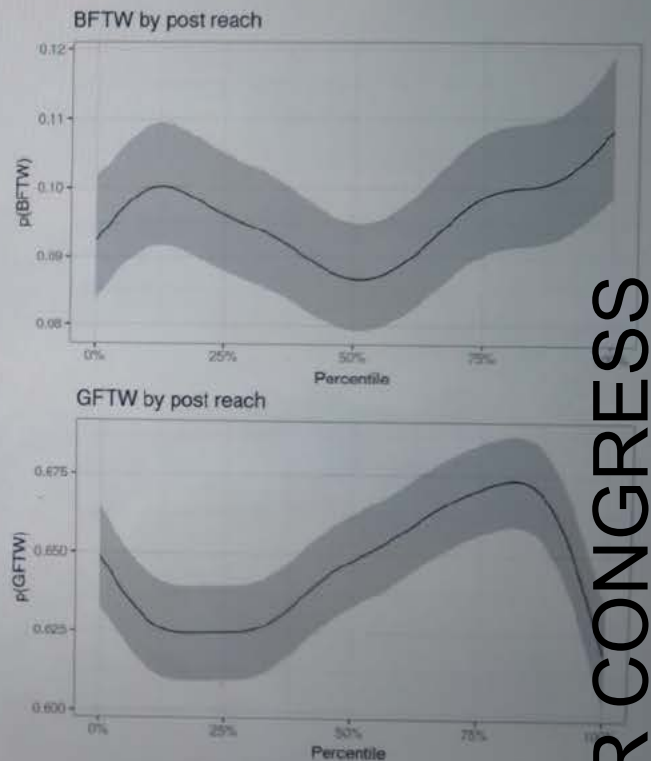
REDACTED FOR CONGRESS

## High-reach posts are more likely to be considered BFTW and less likely to be considered GFTW

These plots compare posts with different levels of reach by percentile (normalized within country).

- For posts with reach in the 25th to 80th percentile, reach is positively associated with the proportion of GFTW responses
- But for posts with very high reach (especially the top 1-2 percentile) GFTW prevalence is lower and BFTW prevalence is higher than average

BFTW = "somewhat" or "very" bad  
GFTW = "somewhat" or "very" good

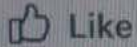


REDACTED FOR CONGRESS

October 23, 2020 · 🌐

High-reach posts are more likely to be considered BFTW and less likely to be considered GFTW ... than lower-reach posts

5 Comments



Like



Comment



Share

I think this headline conclusion is a bit misleading. Even at percentile 100%,  $p(\text{GFTW}) > p(\text{BFTW})$ . I understand that you are drawing a comparison relative to the rest of the distribution for each, so I would recommend reframing the headline using 'disproportionately' or something to that effect.

Like · Reply · 29w

yes the implicit end of that title is "compared with lower reach posts"

Like · Reply · 29w

makes sense. For more context on my comment, I am thinking about the convo in this [thread] ([https://fb.workplace.com/./permalink/3048595641847559/...](https://fb.workplace.com/./permalink/3048595641847559/)) where the same distinction specifically came up.

Like · Reply · 29w

Write a reply...

For posts with reach in the 50th to

Write a comment...





Like · Reply · 29w



██████████ ██████████ makes sense. For more context on my comment, I am thinking about the convo in this [thread] (<https://fb.workplace.com/..../permalink/3048595641847559/...>) where the same distinction specifically came up.

Like · Reply · 29w

Write a reply...  

██████████ ██████████ For posts with reach in the 50th to 75th percentile, reach is positively associated with both  $p(\text{BFTW})$  and  $p(\text{GFTW})$ . Can you explain why this is the pattern? Is it because  $p(\text{BFW}) = 1$  when ratings are 'somewhat/very bad', and  $p(\text{BFW}) = 0$  when ratings are neutral or 'somewhat/very good'.

Clarifying question: is reach 'VPV'?

Like · Reply · 27w · Edited



██████████ ██████████ Essentially the pattern suggests less "neutral" ratings for posts in this range (GFTW = 4,5; BFTW = 1,2; neutral = 3). And yes reach here means total VPVs

Like · Reply · 27w



## Top-N posts are less likely to be considered GFTW and just as likely to be considered BFTW

Top-N posts were less likely to be GFTW than lower-reach posts, odds ratio = .89 ( $p < .001$ )

- Prevalence of BFTW responses did not differ between top-N and lower-reach content
  - However, posts in the top 1 percentile in terms of reach were significantly more likely to be rated BFTW than posts in the bottom 99 percentile (see previous slide)
- Top-N was defined as the top 10,000 posts in terms of unique VPVs across all of Facebook (during the time period the study was conducted)
- These comparisons controlled for user demographic features: age, gender, tenure, friend count, country, and locale

1

9 Comments

Like

Comment

Share

What if we crossed Top N with Civic? Is Top N in Civic more or less likely to be perceived GFTW? This would be a great metric of perception of viral civic content cc

Like · Reply · 29w

1

I think there's definitely something in the idea. I think's concern around aggregate trends harming minority groups in some communities is something to keep in mind, too. I think as we align on XI in how best to measure "good" in the Top - N, we can work together to get a good civic version, too!

Like · Reply · 29w

1

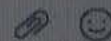
I liked's ideas about creating a classifier based metric that we could be examined more readily for bias and could be constructed with representativeness of minorities in mind. It's also a lot more maintainable.

Like · Reply · 29w

1

- you wouldn't happen to have enough data to do a

Write a comment...



get a good civic version, too!

Like · Reply · 29w



1

I liked [redacted] ideas about creating a classifier based metric that we could be examined more readily for bias and could be constructed with representativeness of minorities in mind. It's also a lot more maintainable.

Like · Reply · 29w



1

[redacted] - you wouldn't happen to have enough data to do a quick replication of this slide solely within civic content? Actually if you had the data, slicing this slide by content category would be interesting.

Like · Reply · 29w

[redacted] Pretty sure we won't have enough data to cross top-N with civic, but I could look at high/low reach x civic (i.e., recreate previous slide w civic)

Like · Reply · 28w

[redacted] - if that's possible, that would be interesting. No need to make a graph and the question I'm trying to answer is whether certain topics have bigger problems with high reach BFTW content than others. My intuition is that the results of the good/bad X reach slide would differ X topic.

Write a comment...



[redacted] - if that's possible, that would be interesting. No need to make a graph and the question I'm trying to answer is whether certain topics have bigger problems with high reach BFTW content than others. My intuition is that the results of the good/bad X reach slide would differ X topic.

Like · Reply · 28w

[redacted] Here is the prevalence of BFTW by topic for "high" (top 2.5%) and "non-high" (bottom 97.5%) reach posts. Looks like your intuition is right there are some interesting differences for high vs lower reach posts across topics (note I switched to SEs here rather than 95% CIs to help compare, and only including cases where there were at least 50 posts matching the category/reach bin)



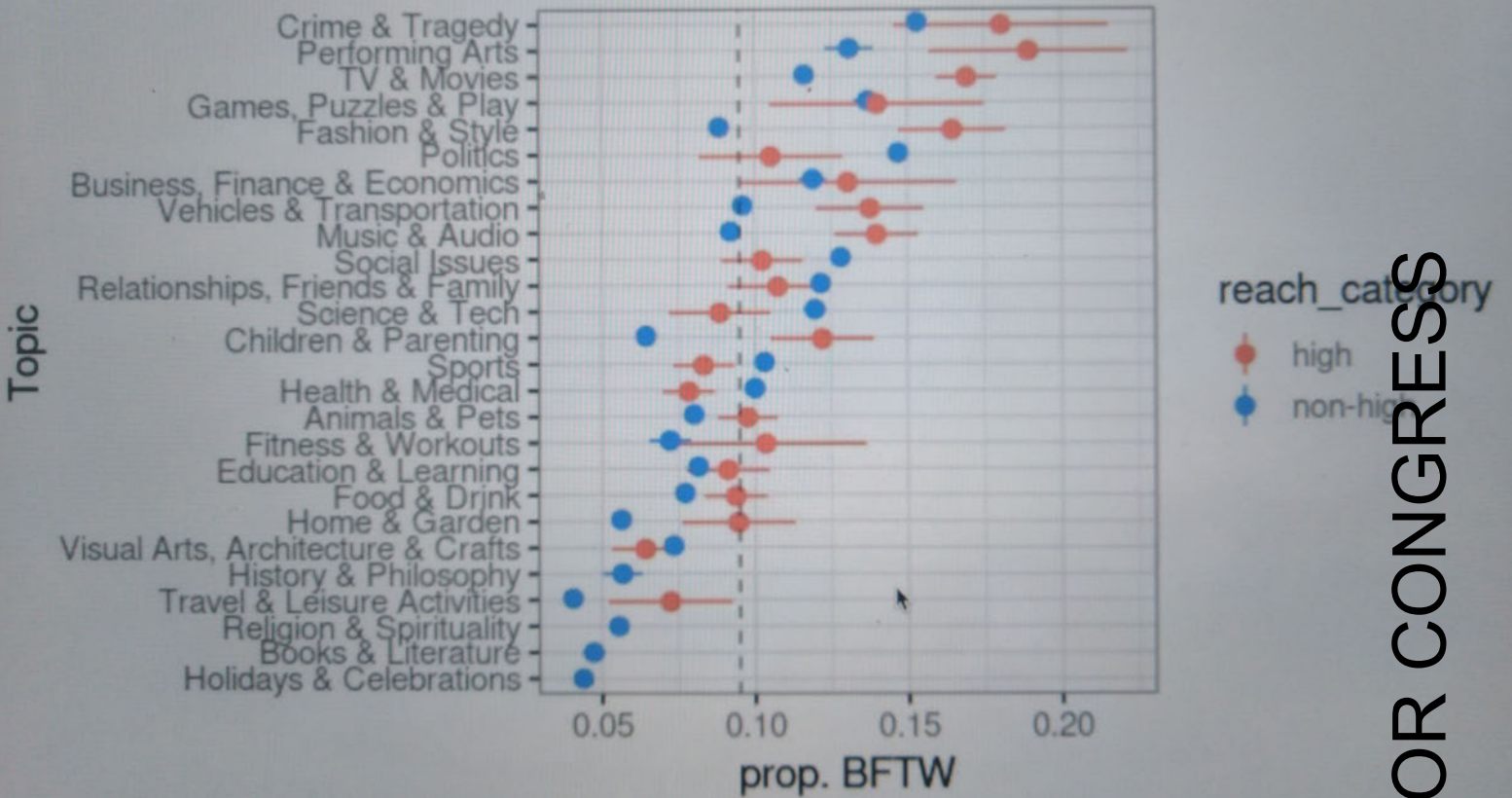
Like · Reply · 28w

[redacted] interesting that politics & social issues (Civic) content actually shows more BFTW for non-high reach content. That's the opposite of my intuition, but then again, if

Save

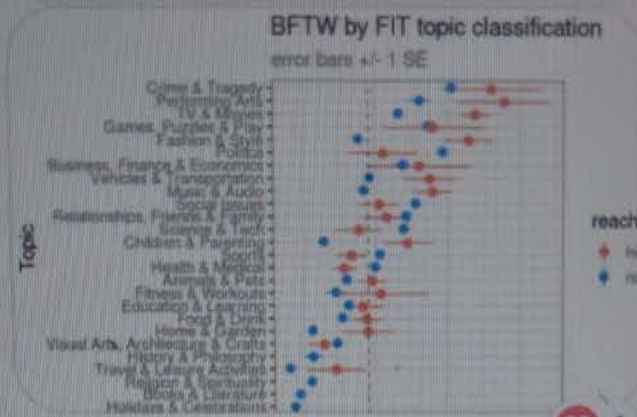
# BFTW by FIT topic classification

error bars +/- 1 SE



REDACTED FOR CONGRESS

at least 50 posts matching the category/reach bin)



Like · Reply · 28w



interesting that politics & social issues (Civic) content actually shows more BFTW for non-high reach content. That's the opposite of my intuition, but then again, if we're surveying people who this is in inventory for, then perhaps that makes sense. Do you have plans to survey random users vs. users who have this in inventory in the future?

Like · Reply · 28w

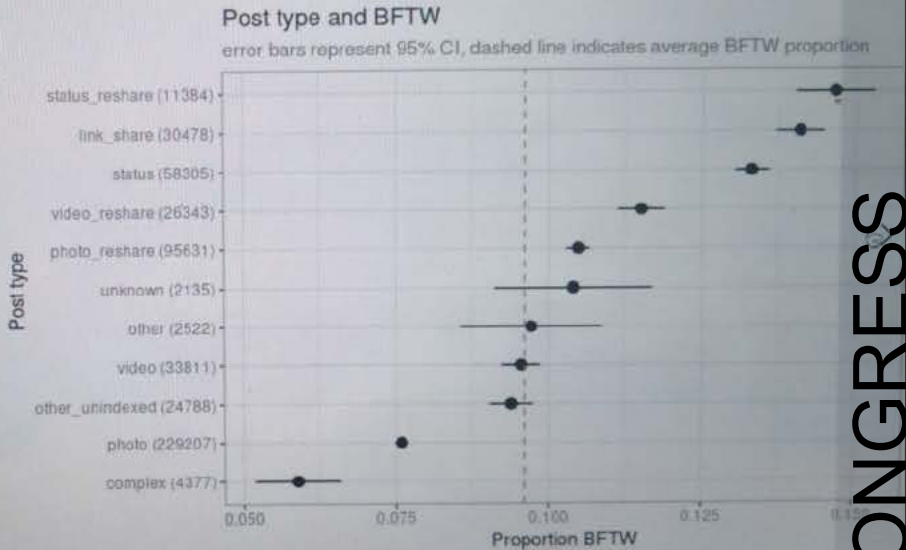


No solid plans yet but there are some conversations about surveys with varying audiences like that



## Reshares, links, and status updates are more likely to be BFTW; photos are less likely to be BFTW

- **Reshares** are an affordance that disproportionately supports distribution of BFTW content
- **Status** posts have the highest prevalence of BFTW among original broadcast post types








REDACTED CONGRESS




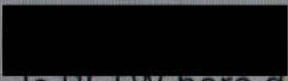
  4

6 Comments



 Like  Comment  Share




Like · Reply · 28w


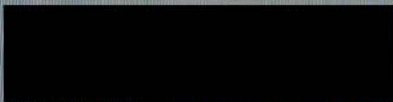
  Is BFTW here defined as selecting one of the two "bad" response options?

Like · Reply · 28w · Edited

  yup




Like · Reply · 28w

 Write a reply...  



  context:  
BFTW means 'bad for the world'

Like · Reply · 27w · Edited




 2




  amazing  1

Like · Reply · 27w

Like · Reply · 27w

 Write a reply...  

 Write a comment...  

portion



0.150

ch 9

## Original status posts were rated BFTW at surprisingly high rates

“Not seeing enough posts from my friends” a common top user pain-point, so it was surprising that users rated original status posts (which typically come from friends) as BFTW more often than many other kinds of content. I have two speculative hypotheses about why this might be:

### Hypothesis #1

- Many users do not share status updates, and those that do are more likely to post negative, divisive, or disrespectful content
  - Prior research has found that many users are discouraged from sharing by disrespectful discourse (e.g., █████ 2020)
  - Users who do choose to share within this ecosystem may be those who are more likely to post BFTW content

### Hypothesis #2

- Users are more likely to post original status posts when they feel very strongly about something
  - Because posting is relatively rare, many users perceive a relatively high cultural/communicative bar to posting
  - They are most likely to exceed that threshold and choose to post when they have strong feelings
  - These charged posts may be seen as BFTW

█████ 2020: <https://fb.workplace.com/photo/> █████  
Credit to █████ for hypothesis 2

October 23, 2020 · 🌐

👍 1

4 Comments



Like



Comment



Share

I have another hypothesis: There are a lot of low-quality copy-and-paste status posts that are effectively reshares, but we wouldn't categorize them as such. I don't know if there are enough of these to explain the high BFTW rate, though.

Like · Reply · 29w

👍 3

Like · Reply · 28w

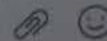
Were these hypothesis arrived at after looking at the actual posts rated BFTW?

Like · Reply · 23w

No this isn't based on a content analysis, more speculation about possible generative processes

Like · Reply · 22w

Write a reply...



Write a comment...

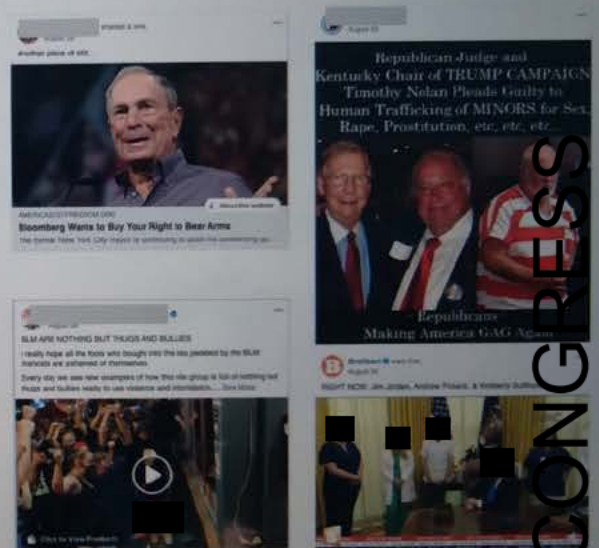


## Civic VPVs are more likely to be considered BFTW

Globally, respondents rated 13.1% CI<sub>95</sub> [12.7% to 13.5%] posts as "very" or "somewhat" bad for the world (unweighted)

- U.S. respondents were nearly 3x as likely to rate civic posts BFTW (13.5%) as non-civic posts (5%)
- Civic posts defined as posts with civic score  $\geq .60$

Sample of civic posts rated BFTW by U.S. respondents



Facebook App Research

REDACTED FOR CONGRESS

15 Comments 2 Shares

👍 Like    💬 Comment    ➦ Share

How representative are these images from the larger spectrum? Are they all this hyperpartisan, or are there more middle line civic ones too? I'm interested to see more data here, because "civic" is pretty broad to describe these, whereas we might be able to much more specifically target this with civic + <whole bunch of other bad signals>...?

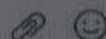
Like · Reply · 29w

These are reasonably representative, especially of posts where there was some agreement the post was BFTW (where it was rated BFTW by more than one respondent)

Like · Reply · 29w

I guess what I'd like to know then are what are the civic posts that are not rated in this way. Because these are definitely "exceptional" when it comes to the larger set of civic content posted and definitely not on the "neutral" or "moderate" side of things. Are there subgroups of civic that are better than others (e.g. news middle-ground

Write a comment...



spondents

ad  
CAMPAIGN  
guilty to  
RS for Sex,  
e, etc...



Again

y Quilloyle...



earch    11

than one respondent)

Like · Reply · 29w

I guess what I'd like to know then are what are the civic posts that are not rated in this way. Because these are definitely "exceptional" when it comes to the larger set of civic content posted and definitely not on the "neutral" or "moderate" side of things. Are there subgroups of civic that are better than others (e.g., news, middle-ground sources, etc.)?

Like · Reply · 29w

Write a reply...



country-by-country breakdown:



Like · Reply · 29w



oh and those errorbars are +/- 1 SE

Like · Reply · 29w

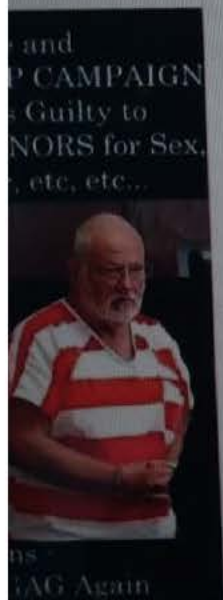
What's the scale of this? Is this 5%-25% or 0.05%-0.25%?

Like · Reply · 29w

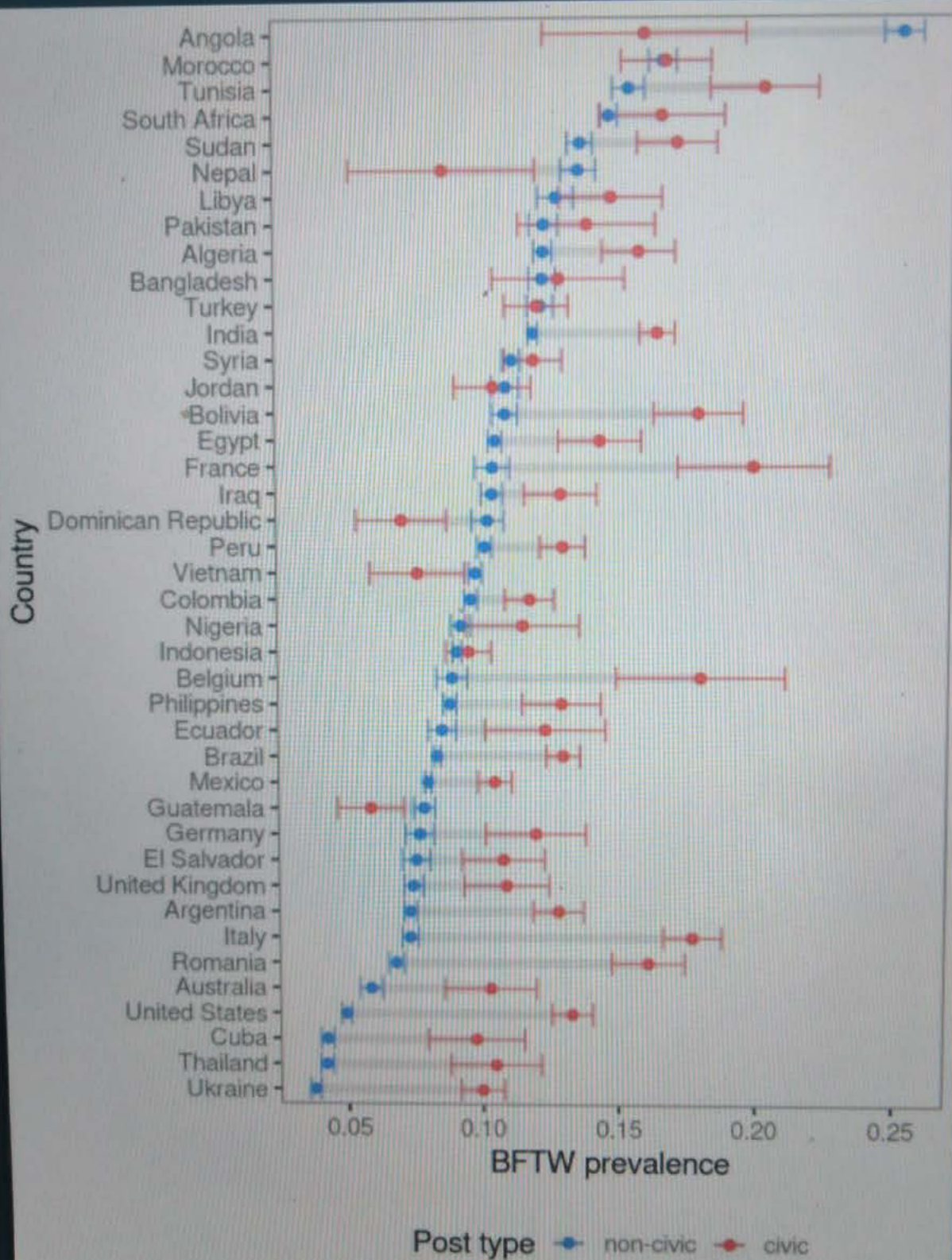
Write a comment...




espondents



Research 11



REDACTED FOR CONGRESS

Like · Reply · 29w  1


oh and those errorbars are +/- 1 SE

Like · Reply · 29w

What's the scale of this? Is this 5%-25% or 0.05%-0.25%?



Like · Reply · 29w

those are proportions so 5-25%

Like · Reply · 29w  1

In a few countries, non-civic is worse than civic. Any hypothesis on what country specific factors are going on in those countries?



Like · Reply · 23w

Write a reply...  

how does civic Good for the world prevalence compare to non-civic GFTW prevalence?

Like · Reply · 29w

the figure above shows this by country, I don't have the overall prevalence in non civic in front of me but it's very similar to overall

Write a comment...  



how does civic Good for the world prevalence compare to non-civic GFTW prevalence?

Like · Reply · 29w

the figure above shows this by country, I don't have the overall prevalence in non civic in front of me but it's very similar to overall average (civic making up a fairly modest slice of all content)

Like · Reply · 29w

maybe i am misreading that chart, but isn't it showing bad for the world? my question is whether good for the world is over- or under-represented in civic content.

Like · Reply · 29w

oh! right sorry I misread.

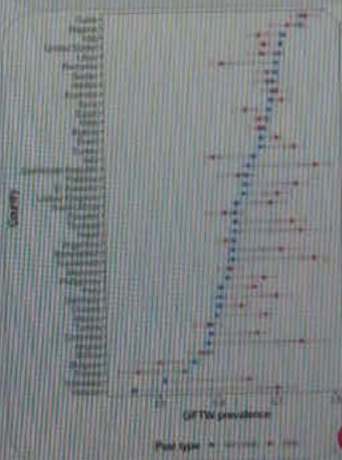
Glad you asked because that turns out to be quite interesting, w/ more heterogeneity across countries. \_\_For a fair number of countries, civic content is more GFTW than non-civic\_\_ (so basically, civic is less likely to be neutral, which makes sense given it is more likely to have real world impact). \_\_The U.S. is an exception here though, civic is both more likely to be BFTW and less likely to be GFTW.\_\_ It's worth noting some of this might be due to

Like · Reply · 29w

[Redacted] oh! right sorry I

misread.

Glad you asked because that turns out to be quite interesting, w/ more heterogeneity across countries. \_\_For a fair number of countries, civic content is more GFTW than non-civic\_\_ (so basically, civic is less likely to be neutral, which makes sense given it is more likely to have real world impact). \_\_The U.S. is an exception here though, civic is both more likely to be BFTW and less likely to be GFTW.\_\_ It's worth noting some of this might be due to language/cultural differences in how respondents used the scale (see slide 20).

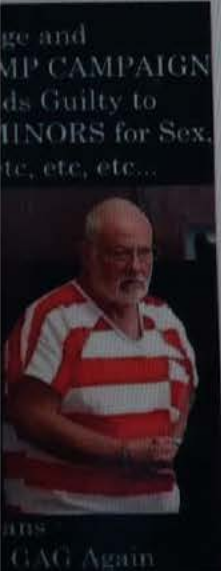


Like · Reply · 29w · Edited

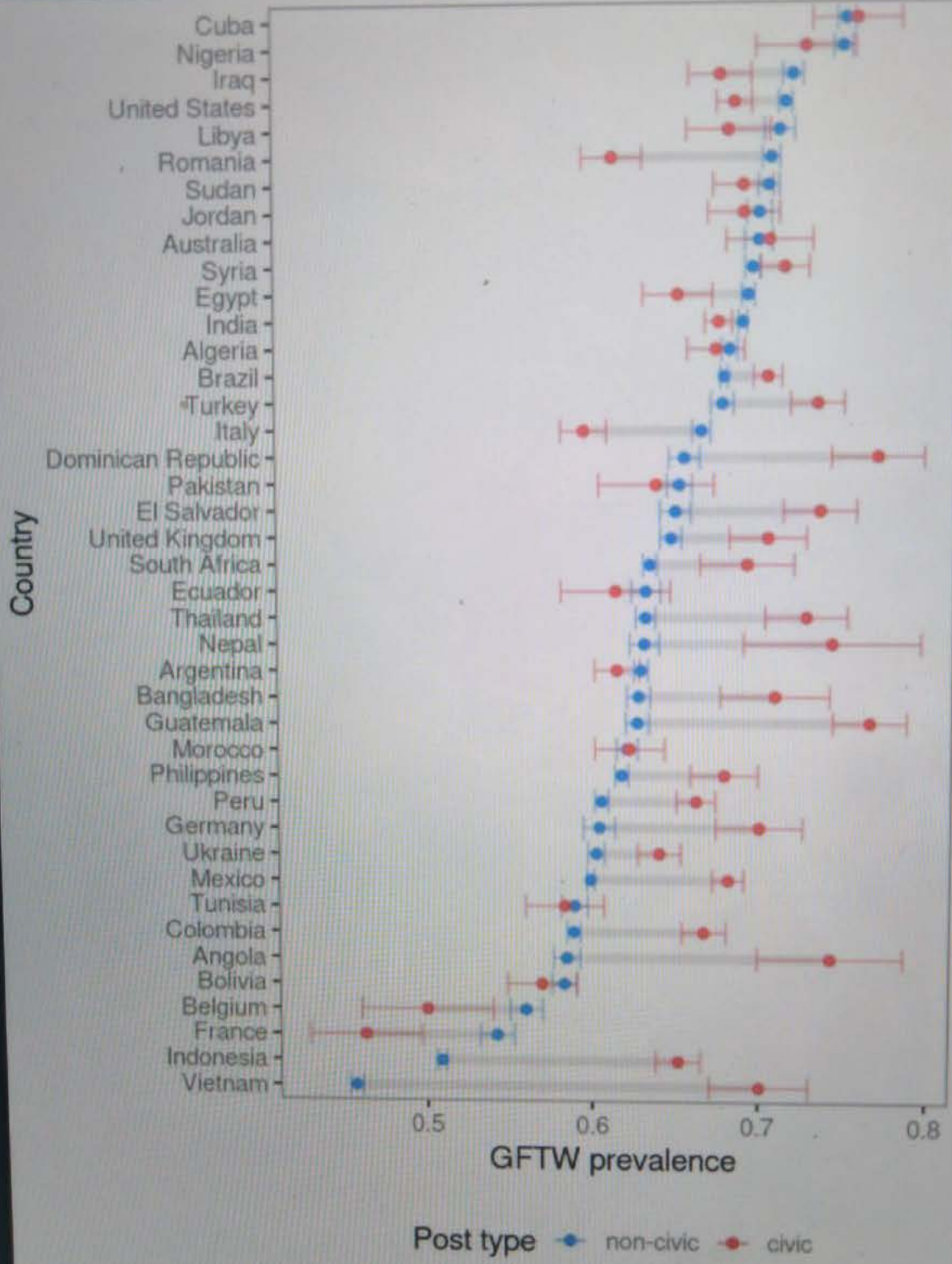
[Redacted] thank you for adding that chart!

Like · Reply · 29w

respondents



Research 11



REDACTED FOR CONGRESS

worth nothing some of this might be due to language/cultural differences in how respondents used the scale (see slide 20).



Like · Reply · 29w · Edited

thank you for adding that chart!

Like · Reply · 29w

Write a reply...

would be interested to know what these prevalence rates would be if they are weighted.

Like · Reply · 28w

These aren't weighted as I wasn't quite sure what the best target population would be (DAP? DAP with civic content exposure? something else?). Open to suggestions though, do you have any thoughts?

Like · Reply · 28w

Write a reply...

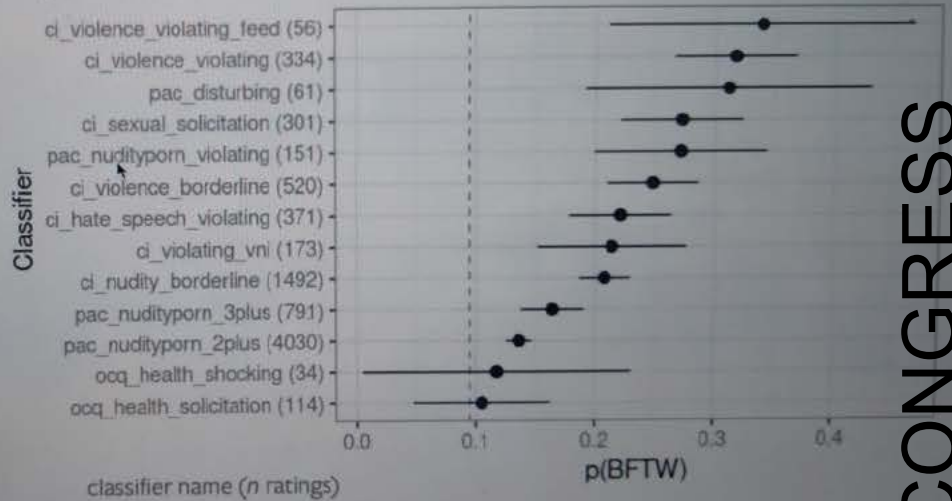
Write a comment...

# VPVs on posts flagged by some integrity classifiers are more likely to be BFTW

BFTW prevalence was higher among posts predicted to contain:

- violence
- sexual solicitation
- hate speech
- violence incitement
- nudity

BFTW by integrity classifier  
Threshold = .60, 95% CIs

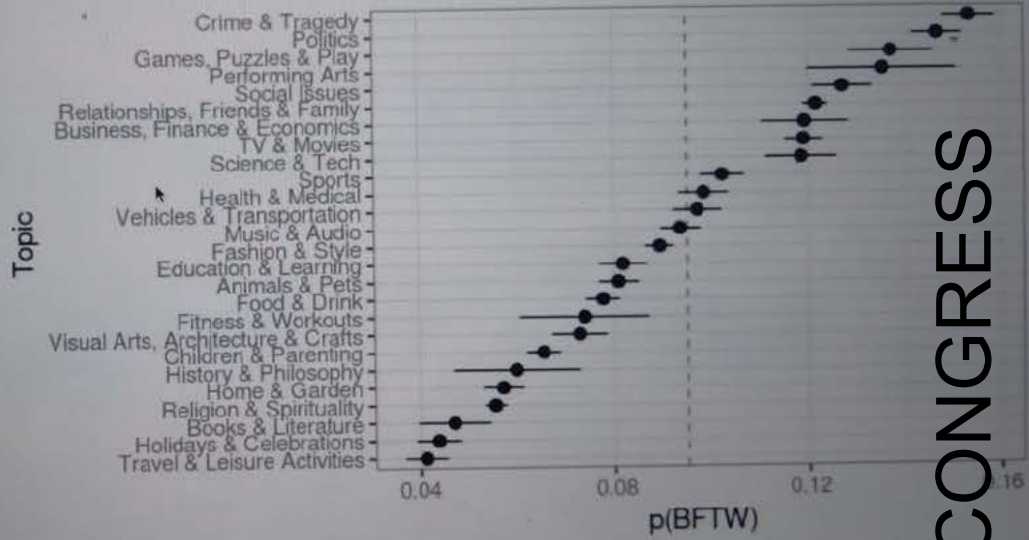


REDACTED FOR CONGRESS

# VPVs on some topics are more likely to be BFTW than others

- It may be surprising that some topics were more likely to be BFTW, such as:
  - Games
  - TV & movies
- But large media such as blockbuster video games and movies can be cultural flashpoints
  - e.g. controversy over gay characters in Star Wars

BFTW by FIT topic classification (Facebook Interest Taxonomy)



REDACTED FOR CONGRESS

Like

Comment

Share

[Redacted] look at #1 topic by bad for the world

Like · Reply · 29w

3

[Redacted]

Like · Reply · 29w

1

[Redacted] - for the Games topic, I'm not seeing a pattern in content (e.g. <https://fburl.com/si/138qdv8f>) Thoughts? If these are correlations with classifier scores rather than actual ratings, perhaps it's picking up on how reactions are being used in that domain?

Like · Reply · 29w

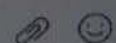
[Redacted] these are the rates of survey responses for content with a topic classifier score above .60 I believe. I guess keep in mind the bftw prevalence is only ~.15 even for high bftw topics ? Otherwise I can pull the fbids for the labeled content if... **See More**

Like · Reply · 28w

1

[Redacted] That makes sense. If it's easy and you have it handy, I'd be happy to use a query that pulls labelled BFTW data X

Write a comment...



rch 13

up on how reactions are being used in that domain?

Like · Reply · 29w

these are the rates of survey responses for content with a topic classifier score above .60 I believe. I guess keep in mind the bftw prevalence is only ~.15 even for high bftw topics ? Otherwise I can pull the fbids for the labeled content if...  
**See More**

Like · Reply · 28w



1

That makes sense. If it's easy and you have it handy, I'd be happy to use a query that pulls labelled BFTW data X category and see if we can generate some hypotheses.

Like · Reply · 28w



1

here's a daiquery query that pulls content fbids and the primary topic tag for each:  
<https://fburl.com/daiquery/ce45xne>

Like · Reply · 28w



1

Write a reply...

re crime and tragedy

Like · Reply · 28w

--relevant to our topic based work for F&F

Like · Reply · 28w



1

Write a comment...



1

0.16



## Greater proportions of VPVs are BFTW in parts of Africa, the Middle East, and Portugal

The color of each country represents the **estimated odds** that users in that country rate posts in their News Feeds BFTW relative to users in the U.S. (OR = 1 means same as U.S.)

Highlighted countries are those where the odds of users rating content BFTW were > 2x as high as in the U.S.

**Caveat:** further investigation is needed to determine whether translation and language differences could partly explain these differences

BFTW relative prevalence  
Odds ratio relative to U.S.



REDACTED FOR CONGRESS

October 23, 2020 · 🌐

👍 2

15 Comments

👍 Like

💬 Comment

➦ Share

Though you might be interested/have insights here

Like · Reply · 28w

^ Hide 14 Replies

I'd love to know what specifically they were labeling, esp things compared to other regions. Have content examples? And what's the fielding time period?

Like · Reply · 28w

Are there samples somewhere? Would be interested to know the proportions of say inflammatory vs low quality vs non-heteronormative.

Like · Reply · 28w

I'm going to put together a more accessible table so folks can pull some of this content and see for themselves, I'll try to get that done this afternoon

Like · Reply · 28w

Great thanks! From

Write a comment...

OR



Like · Reply · 28w



Great thanks! From your familiarity with the content so far, do you have ~-tyrrany~- tyranny of the majority concerns or does "bad for the world" really seem bad for the world, in general.

Like · Reply · 28w · Edited



I definitely have concerns about tyranny of the majority w/ this approach tho they're more theoretical and less about what I've seen in the content per se

Like · Reply · 28w

for anyone coming across this, please see this post with data table:

[https://fb.workplace.com/permalink.php?story\\_fbid=364519941559361&id=100040040738159](https://fb.workplace.com/permalink.php?story_fbid=364519941559361&id=100040040738159)

Like · Reply · 28w



How does GFTW/BFTW consolidate among countries? If my bay area rainbow flag makes its way to uganda, does their survey response affect the US?

Like · Reply · 28w

The "country" variable in the table is the country for the user completing the survey. So if your rainbow flag was rated

Write a comment...



OR



research 14

How does GFTW/BFTW consolidate among countries? If my bay area rainbow flag makes its way to uganda, does their survey response affect the US?

Like · Reply · 28w

The "country" variable in the table is the country for the user completing the survey. So if your rainbow flag was rated by a Ugandan user, then the country variable will be Uganda

Like · Reply · 28w

Cool. It'd be interesting to see country-pair agreement rates on what is good and bad for the world.

Like · Reply · 28w

don't think we can calculate anything like that as these are all ratings of different randomly-sampled content, so there's a fairly small set of content with multiple ratings

Like · Reply · 28w

Lame, lame.

Like · Reply · 28w 🤔 1

How about aggregated at source\_country, rather than individual pieces? Is content

Write a comment...

OR



can calculate anything like that as these are all ratings of different randomly-sampled content, so there's a fairly small set of content with multiple ratings

Like · Reply · 28w

Lame, lame.

Like · Reply · 28w 🤔 1

How about aggregated at source\_country, rather than individual pieces? Is content from the US disproportionately BFTW in Nigeria etc..

Like · Reply · 28w

did you have any hypotheses for this? I think we'd need some idea of what to look for a priori b/c w/ 100-200 countries we're talking about tens of thousands of pairwise comparisons

Like · Reply · 28w

What do I look like, some kind of research guy!? I do not have any strong hypotheses, just generally interested in how different norms work across countries. I guess that's hard to unpick from the selection of content from country x that makes it to country y huh.

Like · Reply · 28w

Write a reply...

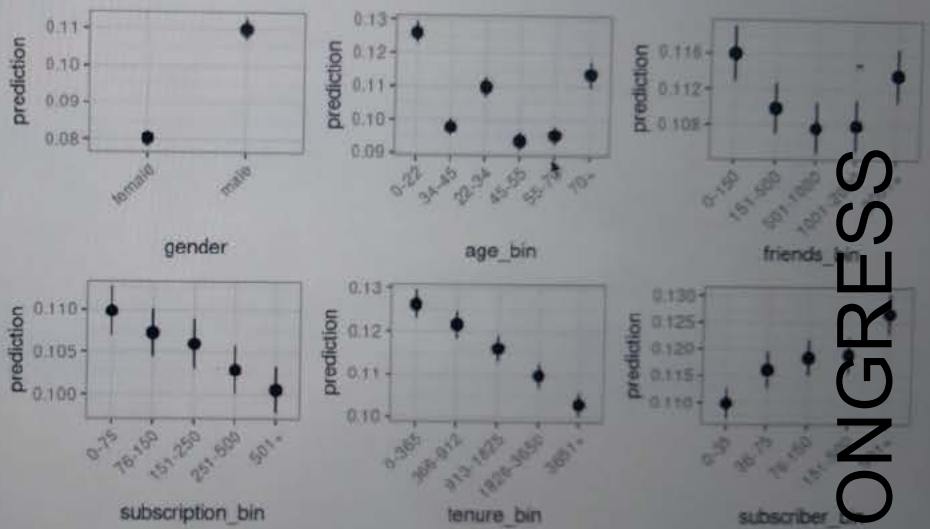
Write a comment...

## The prevalence of BFTW posts in users' feeds varied by respondent demographics

Respondents were more likely to say the posts in their News Feeds were BFTW if they were:

- Men
- From older or younger age groups
- Less tenured
- Users with very few or very many connections

BFTW





REDACTED FOR CONGRESS


October 23, 2020 · 📷

 1

2 Comments

 Like

 Comment

 Share

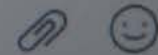
 1

Like · Reply · 29w

- do you have hypotheses as to why these variances based on demos exist?

Like · Reply · 28w

Write a comment...



REDACTED FOR CONGRESS



Save

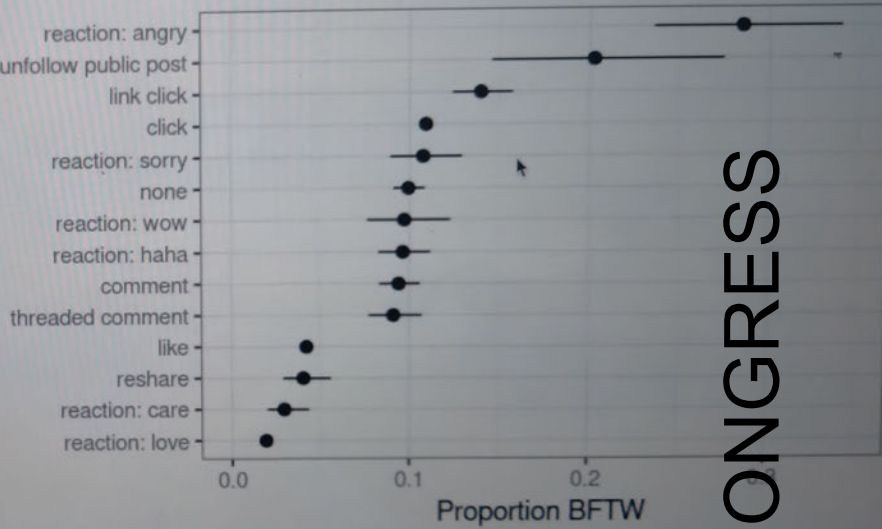
# Engagement data could be used to predict GFTW ratings and to assist in the training of p(GFTW) models

This plot shows the proportion of respondents' ratings for content that they themselves had engaged with (n = 93,626 posts)

- **Angry** reactions were the engagement most positively associated with BFTW ratings
- **Love** reactions were the engagement most negatively associated with BFTW ratings

engagement

BFTW and engagements  
error bars represent 95% CI



Engagements, especially reactions, were correlated with GFTW responses

REDACTED FOR CONGRESS



October 23, 2020 · 🌐

👍 😡 5

9 Comments 1 Share

👍 Like    💬 Comment    ➦ Share

good signal that our Proxy for Anger/Hide/Unfollow is related to BFTW.

Like · Reply · 29w

👍 3

Like · Reply · 28w · Edited

👍 3

This is super interesting. Can I just make sure I'm reading this right.

Angry reactions: this is saying that, of all the posts with angry reactions on them, ~28% were rated as "bad for the world."

Which is the highest - but still implies that 72% were NOT rated BFTW (?)

Like · Reply · 28w · Edited

👍 2

Yup you're reading that exactly right, 72% were rated either "neutral" or "good". Also worth noting there are big CIs here b/c there are only a few hundred "angry" reactions in the dataset

Like · Reply · 28w

👍 2

It's different people leaving

Like · Reply · 28w · Edited

2

Yup you're reading that exactly right, 72% were rated either "neutral" or "good". Also worth noting there are big CIs here b/c there are only a few hundred "angry" reactions in the dataset

Like · Reply · 28w

2

It's different people leaving the angry reacts and taking the survey. EDIT: I was wrong.

Like · Reply · 28w · Edited

not here, this is showing the survey responses for the users who engaged w/ the very content they are being asked about (e.g. asking user X about content Y which they had previously reacted to with "angry")

Like · Reply · 28w

1

ah!

Like · Reply · 28w

Write a reply...

can you interreact comments and reacts or are the data too sparse? In particular, i'd be interested to know whether we see comment\*angry being worse than angry and comment\*love being better than love.

Like · Reply · 28w

1

Write a comment

the angry reacts and taking the survey. EDIT: I was wrong.

Like · Reply · 28w · Edited

not here, this is showing the survey responses for the users who engaged w/ the very content they are being asked about (e.g. asking user X about content Y which they had previously reacted to with "angry")

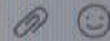
Like · Reply · 28w



ah!

Like · Reply · 28w

Write a reply...



can you interact comments and reacts or are the data too sparse? In particular, i'd be interested to know whether we see comment\*angry being worse than angry and comment\*love being better than love.

Like · Reply · 28w



too sparse unfortunately. One proposal was to run this survey again but upsample posts where users had reacted. We might do that yet tho there isn't a ton of motivation on the product side

Like · Reply · 27w

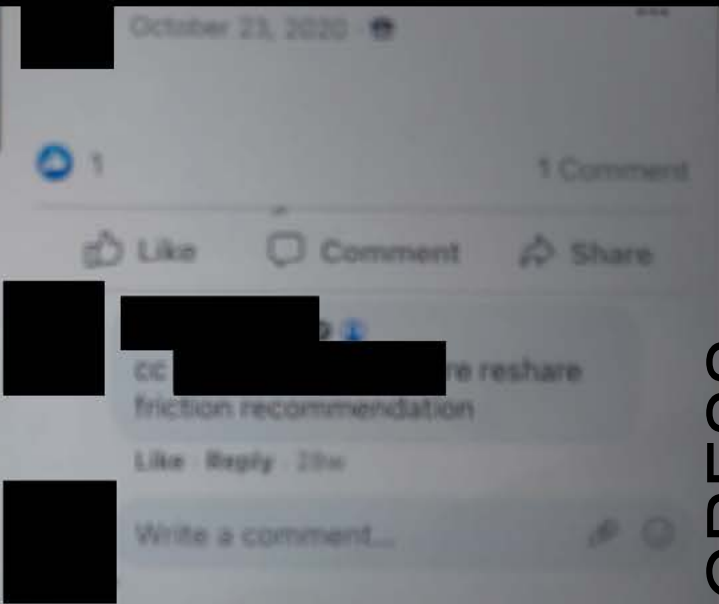
## Recommendations for prioritization

The Good For The World survey gives insight into users' subjective good and bad experiences with content on-platform.

- 1
  - The prevalence of BFTW content in high-reach posts is higher than lower-reach posts
  - We need to do more to discourage the distribution of bad content and foster distribution for good content on the platform.
- 2
  - Higher prevalence of BFTW content in reshares suggests this affordance is disproportionately being used to boost negative content.
  - More work on reshare friction could help reshares more positively impact the content ecosystem
- 3
  - Civic content is more likely to be BFTW than non-civic content
  - Civic and personalization integrity teams should partner to address civic content causing bad experiences for users
  - Guardrails should be developed to ensure efforts to reduce BFTW content do not improperly impact civic speech
- 4
  - There appear to be serious integrity and/or quality gaps in certain countries, especially Portugal and some Middle Eastern and African countries.
  - Additional resources should be directed toward improving the experiences of users in these and other under-served markets

Facebook App Research

REDACTED FOR CONGRESS



experiences with content

4

- There appear to be serious integrity and/or quality gaps in certain countries, especially

REDACTED FOR CONFIDENTIALITY

03  
—  
Can Facebook stand  
behind GFTW content?

REDACTED FOR CONGRESS

## Can Facebook stand behind content if a majority of the users who see that content feel it is GFTW?

One proposed direction for Connection Integrity is that **Facebook should stand behind what we amplify** and that the pillar could achieve this by ensuring that posts gaining large distribution on platform are perceived by a majority of users as GFTW

**This could be very useful, but a few limitations need to be acknowledged or addressed:**

1. The GFTW survey was designed to identify (subjectively) **bad** content, not to provide a true **scale of content quality**
2. GFTW survey responses are **subjective** and users can be divided in their opinions
3. A GFTW survey metric would measure both ecosystem and ranking quality—it is affected both by the **integrity of the content** on the platform and **who views that content**.

## The GFTW survey was designed to identify bad content, not to provide a scale of content quality

Qualitative and quantitative validation studies have shown that a respondents' "bad" ratings truly represent content that they find to be bad experiences, but "good" responses may be less reliable indicators

- Respondents showed a bias to give positive ratings
  - "Bad" responses consistently indicated negative content
  - But "good" responses were more variable: some respondents reserved "good" for really positive content (medical breakthroughs, charitable causes, social progress), but others included anything humorous or about people they liked
- Respondents relationship with the source of a post can be very important in affecting their GFTW responses
  - E.g., "I thought the meme was funny but the person who posted it was quite annoying...my answer probably would have changed if I liked the person ... I gave it 'neutral' " — Participant, age 35

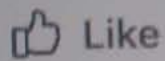
REDACTED FOR CONGRESS



October 23, 2020 · 🌐



2 Comments



Like



Comment



Share

This seems to raise some important questions about use of this question in terms of both validity and reliability.

Like · Reply · 28w

Like · Reply · 28w

Write a comment...



# What users feel is GFTW is highly subjective and frequently does not align with Facebook's integrity and quality standards

These 5 posts were found in a random sample of 25 "very good" ratings from U.S. users



QAnon content



condemnation of a police shooting victim



borderline nudity



violence



low-quality engagement bait

Do we have a sense of low integrity content prevalence among GFTW posts?

Like · Reply · 29w



This isn't super rigorous, but I just quickly looked at what % of rated content scored above .60 on any of the integrity classifiers in the figure on slide 12.

BFTW = 2.3% of content  
Neutral = 1.5% of content  
GFTW = 1.1% of content

Like · Reply · 28w



Worth noting only the borderline nudity example here could have plausibly been caught by those integrity classifiers

Like · Reply · 28w



yeah it's a bit hard to tell b/c 0.6 means different things for different classifiers...

Like · Reply · 28w

yes I'd say just use those as a signal of how prevalence varies across ratings

Like · Reply · 28w

Regardless, idk the CIs but if this is VPV based, integrity - bad content is twice as prevalent in BFTW content



low-quality / engagement bait

.00 on any of the integrity classifiers in the figure on slide 12.

BFTW = 2.3% of content  
Neutral = 1.5% of content  
GFTW = 1.1% of content

Like · Reply · 28w



does

Worth noting only the borderline nudity example here could have plausibly been caught by those integrity classifiers

Like · Reply · 28w



yeah it's a bit hard to tell b/c 0.6 means different things for different classifiers...

Like · Reply · 28w

yes I'd say just use those as a signal of how prevalence varies across ratings

Like · Reply · 28w

Regardless, idk the CIs but if this is VPV based, integrity - bad content is twice as prevalent in BFTW content vs. GFTW. This does seem promising

Like · Reply · 28w



Write a reply...



This is a great example of stuff where we need to go beyond sentiment to find/address/

Like · Reply · 23w



Write a comment...



low-quality / engagement bait

## Among 780 posts rated by 5 or more respondents, 64.5% were considered GFTW by a majority of those respondents

But, **respondents still regularly disagree about these posts**: among the 503 posts with a majority of GFTW ratings, 133 were rated BFTW by at least one respondent

### Details:

- 780 posts were rated by 5 or more respondents
  - These were all high-reach posts (with between 40k to 200m VPVs)
- Of these 780 posts, there were 503 posts where 60% or more of these respondents rated the post as "somewhat" or "very" good for the world (64.5% of posts)

## Posts rated GFTW by a majority of respondents are not universally good and may sometimes create bad experiences

Among the 503 posts with a majority of GFTW ratings, 26% (133) were rated BFTW by at least one respondent.

I examined several dozen posts a majority of respondents felt were GFTW. The majority of these were benign or somewhat positive (see Appendix for examples).

However, others could be bad experiences for some users, such as:

- A video compilation of crocodiles eating other animals
- An instance of borderline nudity (PAC 2)
- Political and religious statements (including anti-semitic comments)
- A clickbait-y article

Examples



REDACTED FOR CONGRESS

## Even if the majority of users who see a piece of content on Facebook feel it is good, the majority of all users and external stakeholders might feel it is bad

### Good For The World lies in the eye of the beholder ...

- Because GFTW is a subjective, personalized survey, Facebook may not be willing to stand behind all content a majority of viewers feel is GFTW
  - For instance: An inflammatory post dehumanizing muslims in India might be rated majority GFTW if the majority of users connected to that post hold negative views toward Muslims
- A GFTW survey metric would provide limited insight into ecosystem health
  - Personalized ranking changes could increase the proportion of content rated GFTW without improving the health of the content ecosystem
  - We may simply be showing "bad" content only to those who like to engage with it (increasing filter bubbles)
- Increasing the percent of users who said Top-N content in their Feed is "good" may not impact overall quality or legitimacy of Top-N content
  - The proportion of top-N content surveyed users said was GFTW might be improved even if the Top-N content itself doesn't change

## Is 9.5% bad bad? Is 63.6% good good?

### Absolute prevalence rates can be hard to interpret and evaluate

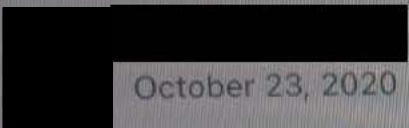
- **We don't really want 100% GFTW content**
  - Some "neutral" content might just be relatively frivolous, yet still positive for users
- **We don't even really want 0% BFTW content**
  - Responses are *subjective*, so even in the most healthy ecosystem and society there will be things that *some* people think are bad for the world.
  - Content users feel is BFTW can sometimes be important or essential content that they need to be aware of
  - 0% BFTW content could be indicative of ranking that creates extremely strong filter bubbles rather than true integrity successes

### Relative comparisons of prevalence of GFTW/BFTW can be more useful

For instance, we can examine:

- **Reach (Top-N):** Is the content winning the greatest distribution more or less good for the world than content with less reach? (see slides 7-8)
- **Reshares:** Does this mechanism build distribution for good or bad content? (see slide 9)
- **User demographics / regions:** Are some kinds of users seeing content that is better or worse than others? (see slides 14-15)





October 23, 2020 · 📍

👍 2

1 Comment

👍 Like    💬 Comment    ➦ Share

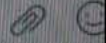


this slide helps address your question from the PL review about relative vs. absolute goaling that we never got back to answering

Like · Reply · 29w



Write a comment...



t  
ontent

for

f users

## Recommendations for metrics and goaling

As a metric, the GFTW survey and model predictions could give valuable insight into the quality of users' News Feed experience and/or the progress of personalized integrity efforts.

1

- BFTW responses more clearly identify "bad" content than GFTW responses identify "good" content
- Majority GFTW content can be divisive and have higher-than-average BFTW scores as well
- Defining a majority-GFTW metric as "neutral or above" could address both of these issues

2

- It could be simpler to set goals based on the relative proportion of GFTW content rather than absolute proportions of GFTW content.
- For instance: we could set a goal that the content winning the greatest distribution on platform (top-N content) ought to be more good and/or less bad for the world than content with less reach.

3

- If a GFTW metric is used to measure content that Facebook would stand behind, the metric must account for the *audience* of posts.
- For instance: a model-based metric could evaluate consensus GFTW for content by computing p(GFTW) scores for a diverse and representative set of users, rather than only those who are connected to that content.

October 23, 2020 · 🌐

👍 2

8 Comments 2 Shares



Like



Comment



Share

I like the idea of a model based metric to extract the positive factor from the subjectivity, to make it more predictable and also from a cost/maintenance perspective.

Like · Reply · 29w

3 is essential; that's one thing I'd like us to push for

Like · Reply · 29w

👍 2

I think all of these insights / recommendations are predicated on the assumption that this is a valid and reliable question when it seems like there is plenty of evidence here to suggest otherwise and that we really don't know what it is measuring. Before discussing use as a metric much less goaling, seems like we need a much clearer sense of what it is exactly that we are trying to measure, and then do the work to design survey question(s) to measure that concept.

Like · Reply · 28w

Hmm I'm not trying to claim that the question is perfect but I think there has been quite a lot of thought put into vetting it. You say "there is

Write a comment...



Hmm I'm not trying to claim that the question is perfect but I think there has been quite a lot of thought put into vetting it. You say "there is plenty of evidence here to suggest otherwise and that we really don't know what it is measuring". I'm not sure I agree with that or didn't read that when I looked through the deck, but can you be more specific about what this evidence is?

I'm not trying to prevent people from providing valid criticism but I do want to make sure people understand the broader context and that generally we push toward solutions. We 100% don't want to launch things that makes things worse or use metrics that are fundamentally flawed. But we do face considerable constraints and are trying to navigate those while still making things better. ...

GFTW / BFTW are not perfect measures but I would argue they push us to a better place than the current system which strictly considers engagement. And if we use these signals in ranking or as metrics they represent a part of the whole, not the only way we evaluate the ecosystem. I'm just not convinced that a world in which we start to think about GFTW is worse than the status quo and that's the most fundamental

Write a comment...

making things better.

GFTW / BFTW are not perfect measures but I would argue they push us to a better place than the current system which strictly considers engagement. And if we use these signals in ranking or as metrics they represent a part of the whole, not the only way we evaluate the ecosystem. I'm just not convinced that a world in which we start to think about GFTW is worse than the status quo and that's the most fundamental question to me.

Like · Reply · 28w

Slide #20 provides several reasons to be concerned about the measurement properties of this item, beyond (IMO) clear face validity concerns. There is also a mismatch between the question stem and the response scale. Beyond these details, I still have not been able to identify what the attitudinal construct that this is supposed to be measuring is. These concerns raise questions about both the validity and reliability of the measure.

I'm not making an argument that something like this might or might not make things better relative to status quo. I'm all in favor of moving beyond an overemphasis on engagement signals (very much so, in fact). It's just not a very high bar to be holding

Write a comment...

response scale. Beyond these details, I still have not been able to identify what the attitudinal construct that this is supposed to be measuring is. These concerns raise questions about both the validity and reliability of the measure.

I'm not making an argument that something like this might or might not make things better relative to status quo. I'm all in favor of moving beyond an overemphasis on engagement signals (very much so, in fact). It's just not a very high bar to be holding ourselves to. Are there reasons why we can't do better or why we are forced to use this "good for world" phrasing?

Like · Reply · 28w

I can't answer your questions around the validity of the measure. I think [redacted] and [redacted] the UXR who worked on this, are pretty thoughtful and likely have considered many of the concerns folks have raised. They are working within a narrow lane where we are trying to 1) take a personalized approach to Integrity due to policy constraints and 2) achieve broad impact with a single survey / model. It would not shock me if they share some of your concerns, but I imagine we within XI all think this is better than the status quo and thus believe in pushing it forward. I'll also

Write a comment...



I can't answer your questions around the validity of the measure. I think [redacted] and [redacted] the UXR who worked on this, are pretty thoughtful and likely have considered many of the concerns folks have raised. They are working within a narrow lane where we are trying to 1) take a personalized approach to Integrity due to policy constraints and 2) achieve broad impact with a single survey / model. It would not shock me if they share some of your concerns, but I imagine we within XI all think this is better than the status quo and thus believe in pushing it forward. I'll also just reiterate that this signal and survey don't have to stand on their own; they can be used with other signals and tweaked to try to correct any issues.

Like · Reply · 28w

[redacted] My take from slide 20 was that there are some concerns with trying to use GFTW as a raw signal to identify good content. That said, I don't read it as saying there is no way to make this signal useful for identifying good content, just that on its own there may be some limitations. For instance, we could combine it with p(bad) classifiers or ensure that the GFTW score is high for a diverse swatch of users.

Write a comment...



pushing it forward. I'll also just reiterate that this signal and survey don't have to stand on their own; they can be used with other signals and tweaked to try to correct any issues.

Like · Reply · 28w

My take from slide 20 was that there are some concerns with trying to use GFTW as a raw signal to identify good content. That said, I don't read it as saying there is no way to make this signal useful for identifying good content, just that on its own there may be some limitations. For instance, we could combine it with  $p(\text{bad})$  classifiers or ensure that the GFTW score is high for a diverse swatch of users.

The main purpose of the survey and corresponding model were actually only to identify content that is bad for the world btw. There have been questions about whether we can use it for good for the world content as well, hence some of that debate. I read slide 20 as saying the BFTW approach is acceptable.

Like · Reply · 28w

What are your face validity concerns?

Like · Reply · 28w

Write a reply...

Write a comment...



Save

3

04

# Appendix

27

REDACTED FOR CONGRESS

Waiting for content...

# Majority-GFTW content examples



UNICEF shared a post  
 School reopening must be a priority... but only if they're safe. Handwashing facilities are crucial to protecting children from COVID-19 and other diseases. Even before the pandemic, 2 in 3 schools globally lacked basic handwashing facilities. To reopen schools safely, we have to both our school leaders hand hygiene.



UNICEF shared a post  
 The COVID-19 pandemic has forced remote work into the new norm. And according to a new survey, many older adults think a long-term switch to working from home may just be the perfect fit for their lives. Here's why.



Turn down for what? (I major didn't get what they're saying)



UNICEF shared a post  
 The demand for mental health services has increased significantly in the context of Lebanon due to the impact of a deep economic crisis and the COVID-19 pandemic. UNICEF is working with partners to provide children with psychological support. Contact today to help families rebuild their lives.



Facebook App

These examples were drawn to illustrate the range of positive content marked GFTW based on what I saw in a random sample of about 30 posts. However, these are not perfectly representative as I have selected some of the more "interesting" examples for this slide.

REDACTED FOR CONGRESS

## GFTW responses are distinct from “worth your time” and other content-level metrics

- GFTW responses and p(WYT) model predictions only weakly correlated (Spearman’s rho = .22)
- GFTW responses are more strongly correlated with integrity classifier scores than are other content-level surveys (██████████ 2019; ██████████ 2020)
- GFTW survey responses were not substantially correlated with predicted MSI (rho = .055) nor observed MSI (rho = .063)

Also see ██████████ 2019: <https://fb.workplace.com/photo?fbid=490257215241967&set=a.490255551908800>

REDACTED FOR CONGRESS

## GFTW survey responses are more strongly associated with integrity problems than are Worth Your Time responses

- Compared to responses to WYT and LUV surveys, GFTW responses differ more strongly for posts with:
  - bullying
  - borderline nudity
  - spam
  - 3PFC misinformation
  - borderline hate

Plot displays regression coefficients indicating difference in ratings of classified and benign content.

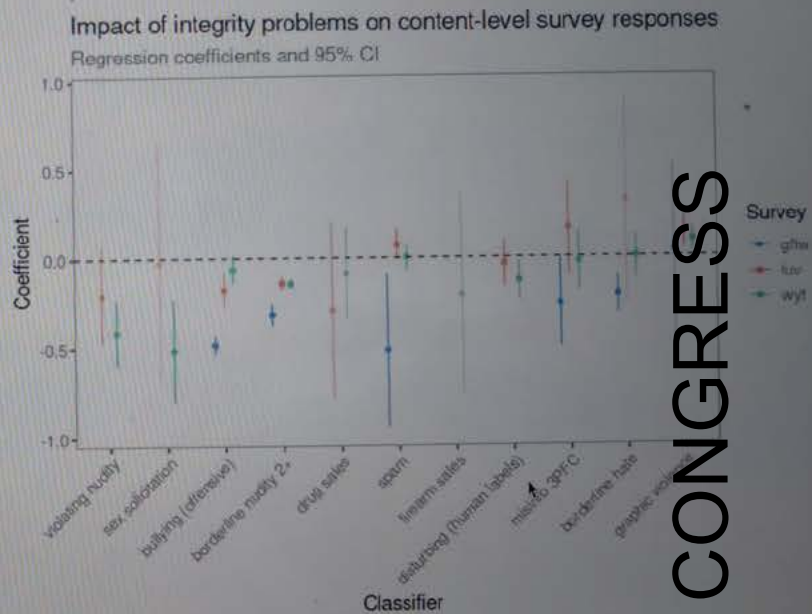


Figure reproduced from [redacted] 2020  
[https://fb.workplace.com/permalink.php?story\\_fbid=269573991053957&id=100040040738159](https://fb.workplace.com/permalink.php?story_fbid=269573991053957&id=100040040738159)

REDACTED FOR CONGRESS