TEC.010

5.1.2.e

Van: 5.1.2.e
Verzonden: woensdag 3 juni 2020 09:55
Aan: 5.1.2.e
Onderwerp: FW: info FB nav misinfo working group call

Van: 5.1.2.e
Verzonden: woensdag 3 juni 2020 09:24
Aan: 5.1.2.e
Onderwerp: info FB nav misinfo working group call

Ha 5.1.2.e,

Zoals beloofd, ter verspreiding in de desinfo groep:

- De tijdslijn van de EOB (welke beslissingen eerst, welke zaken later pas, wat gebeurt er precies na een beslissing, etc.): https://about.fb.com/news/2020/01/facebooks-oversight-board/ Meer info via https://www.oversightboard.com/
- De link naar Mark Zuckerberg zijn post over het al dan niet weghalen van content van president Trump: https://www.facebook.com/zuck/posts/10111961824369871 (tevens hieronder ingeplakt voor wie geen Facebook heeft, eerste bericht)
- De link naar Mark Zuckberg zijn notitie uit 2018 over content governance die eea in perspectief plaatst. NB. Dit stuk is relatief oud, maar geeft een goed idee van de bredere context van onze verantwoordelijkheid mbt content: https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634 (tevens helemaal onderaan geplakt, is vrij lang). Wat betreft de aanpak van misinfo geldt onze bredere strategie: Remove (nepaccounts, overtredingen van de Community Standards), Reduce (verminder verspreiding van materieel dat is gefactcheckt en sensationalism & clickbait), Inform (mediawijsheid stimuleren, context bieden en transparentie)
- Het Twitterdraadje mbt bots en of deze nu wel of niet omni-present zijn: https://twitter.com/5.1.2.e /status/1264956411785641986 (incl. comments van Twitter zijn security lead, onze 5.1.2.e , etc.).

We hebben contact!

Met vriendelijke groet,

5.1.2.e

5.1.2.e

5.1.2.e

Jollemanhof 15 | 1019 GW Amsterdam
Facebook | Mobile 5.1.2.e

Mark Zuckerberg

30 mei om 01:19 ·

This has been an incredibly tough week after a string of tough weeks. The killing of George Floyd showed yet again that for Black people in America, just existing means risking your life. This comes weeks after the killing of Ahmaud Arbery and Breonna Taylor, and in the midst of Covid having a disproportionate impact on the Black community in the US. It continues a long and devastating history of human loss going back centuries. I know the conversations happening amongst our Black friends, colleagues and neighbors are incredibly painful. As Americans, this affects all of us and we all have an obligation to help address the inequality in how justice is served. This is something I care deeply about.

I've been struggling with how to respond to the President's tweets and posts all day. Personally, I have a visceral negative reaction to this kind of divisive and inflammatory rhetoric. This moment calls for unity and calmness, and we need empathy for the people and communities who are hurting. We need to come together as a country to pursue justice and break this cycle.

But I'm responsible for reacting not just in my personal capacity but as the leader of an institution committed to free expression. I know many people are upset that we've left the President's posts up, but our position is that we should enable as much expression as possible unless it will cause imminent risk of specific harms or dangers spelled out in clear policies. We looked very closely at the post that discussed the protests in Minnesota to evaluate whether it violated our policies. Although the post had a troubling historical reference, we decided to leave it up because the National Guard references meant we read it as a warning about state action, and we think people need to know if the government is planning to deploy force. Our policy around incitement of violence allows discussion around state use of force, although I think today's situation raises important questions about what potential limits of that discussion should be. The President later posted again, saying that the original post was warning about the possibility that looting could lead to violence. We decided that this post, which explicitly discouraged violence, also does not violate our policies and is important for people to see. Unlike Twitter, we do not have a policy of putting a warning in front of posts that may incite violence because we believe that if a post incites violence, it should be removed regardless of whether it is newsworthy, even if it comes from a politician. We have been in touch with the White House today to explain these policies as well.

There are heated debates about how we apply our policies during moments like this. I know people are frustrated when we take a long time to make these decisions. These are difficult decisions and, just like today, the content we leave up I often find deeply offensive. We try to think through all the consequences, and we keep our policies under constant review because the context is always evolving. People can agree or disagree on where we should draw the line, but I hope they understand our overall philosophy is that it is better to have this discussion out in the open, especially when the stakes are so high. I disagree strongly with how the President spoke about this, but I believe people should be able to see this for themselves, because ultimately accountability for those in positions of power can only happen when their speech is scrutinized out in the open.

A Blueprint for Content Governance and Enforcement

MARK ZUCKERBERG·DONDERDAG 15 NOVEMBER 2018·LEESTIJD: 22 MINUTEN

My focus in 2018 has been addressing the most important issues facing Facebook. As the year wraps up, I'm writing a series of notes about these challenges and the progress we've made. The first note was about Preparing for Elections and this is the second in the series.

•••

Many of us got into technology because we believe it can be a democratizing force for putting power in people's hands. I've always cared about this and that's why the first words of our mission have always been "give people the power". I believe the world is better when more people have a voice to share their experiences, and when traditional gatekeepers like governments and media companies don't control what ideas can be expressed.

At the same time, we have a responsibility to keep people safe on our services -- whether from terrorism, bullying, or other threats. We also have a broader social responsibility to help bring people closer together -- against polarization and extremism. The past two years have shown that without sufficient safeguards, people will misuse these tools to interfere in elections, spread misinformation, and incite violence. One of the most painful lessons I've learned is that when you connect two billion people, you will see all the beauty and ugliness of humanity.

An important question we face is how to balance the ideal of giving everyone a voice with the realities of keeping people safe and bringing people together. What should be the limits to what people can express? What content should be distributed and what should be blocked? Who should decide these policies and make enforcement decisions? Who should hold those people accountable?

As with many of the biggest challenges we face, there isn't broad agreement on the right approach, and thoughtful people come to very different conclusions on what are acceptable tradeoffs. To make this even harder, cultural norms vary widely in different countries, and are shifting rapidly.

I have focused more on these content governance and enforcement issues than any others over the past couple of years. While it has taken time to understand the complexity of the challenges, we have made a lot of progress. Still, we have significant work ahead to get all our systems to the levels people expect, and where we need to be operating.

Even then, there will always be issues. These are not problems you fix, but issues where you continually improve. Just as a free society will always have crime and our expectation of government is not to eliminate all crime but to effectively manage and reduce it, our community will also always face its share of abuse. Our job is to keep the misuse low, consistently improve over time, and stay ahead of new threats.

In this note, I will outline the approach we're taking. A full system requires addressing both governance and enforcement. I will discuss how we're proactively enforcing our policies to remove more harmful content, preventing borderline content from spreading, giving people more control of their experience, and creating independent oversight and transparency into our systems.

Community Standards

Before getting into what we need to improve, it's important to understand how we've approached these problems until now. Every community has standards, and since our earliest days we've also had our Community Standards -- the rules that determine what content stays up and what comes down on Facebook. Our goal is to err on the side of giving people a voice while preventing real world harm and ensuring that people feel safe in our community. You can read them here: http://www.facebook.com/communitystandards

In April, we went a step further and published our internal guidelines that our teams use to enforce these standards. These guidelines are designed to reduce subjectivity and ensure that decisions made by reviewers are as consistent as possible. For example, our Community Standards on violence and graphic content say "we remove content that glorifies violence or celebrates the suffering or humiliation of others". Sometimes there are reasons to share this kind of troubling content, like to draw attention to human rights abuses or as a news organization covering important events. But there have to be limits, and our guidelines include 18 specific types of content we remove, including visible internal organs and charred or burning people.

The team responsible for setting these policies is global -- based in more than 10 offices across six countries to reflect the different cultural norms of our community. Many of them have devoted their careers to issues like child safety, hate speech, and terrorism, including as human rights lawyers or criminal prosecutors.

Our policy process involves regularly getting input from outside experts and organizations to ensure we understand the different perspectives that exist on free expression and safety, as well as the impacts of our policies on different communities globally. Every few weeks, the team runs a meeting to discuss potential changes to our policies based on new research or data. For each change the team gets outside input -- and we've also invited academics and journalists to join this meeting to understand this process. Starting today, we will also publish minutes of these meetings to increase transparency and accountability.

The team responsible for enforcing these policies is made up of around 30,000 people, including content reviewers who speak almost every language widely used in the world. We have offices in many time zones to ensure we can respond to reports quickly. We invest heavily in training and support for every person and team. In total, they review more than two million pieces of content every day. We issue a transparency report with a more detailed breakdown of the content we take down.

For most of our history, the content review process has been very reactive and manual -- with people reporting content they have found problematic, and then our team reviewing that content. This approach has enabled us to remove a lot of harmful content, but it has major limits in that we can't remove harmful content before people see it, or that people do not report.

Accuracy is also an important issue. Our reviewers work hard to enforce our policies, but many of the judgements require nuance and exceptions. For example, our Community Standards prohibit most nudity, but we make an exception for imagery that is historically significant. We don't allow the sale of regulated goods like firearms, but it can be hard to distinguish those from images of paintball or toy guns. As you get into hate speech and bullying, linguistic nuances get even harder -- like understanding when someone is condemning a racial slur as opposed to using it to attack others. On top of these issues, while computers are consistent at highly repetitive tasks, people are not always as consistent in their judgements.

The vast majority of mistakes we make are due to errors enforcing the nuances of our policies rather than disagreements about what those policies should actually be. Today, depending on the type of content, our review teams make the wrong call in more than 1 out of every 10 cases. Reducing these errors is one of our most important priorities. To do this, in the last few years we have significantly ramped up our efforts to proactively enforce our policies using a combination of artificial intelligence doing the most repetitive work, and a much larger team of people focused on the more nuanced cases. It's important to remember though that given the size of our community, even if we were able to reduce errors to 1 in 100, that would still be a very large number of mistakes.

Proactively Identifying Harmful Content

The single most important improvement in enforcing our policies is using artificial intelligence to proactively report potentially problematic content to our team of reviewers, and in some cases to take action on the content automatically as well.

This approach helps us identify and remove a much larger percent of the harmful content -- and we can often remove it faster, before anyone even sees it rather than waiting until it has been reported.

Moving from reactive to proactive handling of content at scale has only started to become possible recently because of advances in artificial intelligence -- and because of the multi-billion dollar annual investments we can now fund. To be clear, the state of the art in AI is still not sufficient to handle these challenges on its own. So we use computers for what they're good at -- making basic judgements on large amounts of content quickly -- and we rely on people for making more complex and nuanced judgements that require deeper expertise.

In training our AI systems, we've generally prioritized proactively detecting content related to the most real world harm. For example, we prioritized removing terrorist content -- and now 99% of the terrorist content we remove is flagged by our systems before anyone on our services reports it

to us. We currently have a team of more than 200 people working on counter-terrorism specifically.

Another category we prioritized was self harm. After someone tragically live-streamed their suicide, we trained our systems to flag content that suggested a risk -- in this case so we could get the person help. We built a team of thousands of people around the world so we could respond to these flags usually within minutes. In the last year, we've helped first responders quickly reach around 3,500 people globally who needed help.

Some categories of harmful content are easier for AI to identify, and in others it takes more time to train our systems. For example, visual problems, like identifying nudity, are often easier than nuanced linguistic challenges, like hate speech. Our systems already proactively identify 96% of the nudity we take down, up from just close to zero a few years ago. We are also making progress on hate speech, now with 52% identified proactively. This work will require further advances in technology as well as hiring more language experts to get to the levels we need.

In the past year, we have prioritized identifying people and content related to spreading hate in countries with crises like Myanmar. We were too slow to get started here, but in the third quarter of 2018, we proactively identified about 63% of the hate speech we removed in Myanmar, up from just 13% in the last quarter of 2017. This is the result of investments we've made in both technology and people. By the end of this year, we will have at least 100 Burmese language experts reviewing content.

In my note about our efforts Preparing for Elections, I discussed our work fighting misinformation. This includes proactively identifying fake accounts, which are the source of much of the spam, misinformation, and coordinated information campaigns. This approach works across all our services, including encrypted services like WhatsApp, because it focuses on patterns of activity rather than the content itself. In the last two quarters, we have removed more than 1.5 billion fake accounts.

Over the course of our three-year roadmap through the end of 2019, we expect to have trained our systems to proactively detect the vast majority of problematic content. And while we will never be perfect, we expect to continue improving and we will report on our progress in our transparency and enforcement reports.

It's important to note that proactive enforcement doesn't change any of the policies around what content should stay up and what should come down. That is still determined by our Community Standards. Proactive enforcement simply helps us remove more harmful content, faster. Some of the other improvements we're making will affect which types of content we take action against, and we'll discuss that next.

Discouraging Borderline Content

One of the biggest issues social networks face is that, when left unchecked, people will engage disproportionately with more sensationalist and provocative content. This is not a new phenomenon. It is widespread on cable news today and has been a staple of tabloids for more than a century. At scale it can undermine the quality of public discourse and lead to polarization. In our case, it can also degrade the quality of our services.

Our research suggests that no matter where we draw the lines for what is allowed, as a piece of content gets close to that line, people will engage with it more on average -- even when they tell us afterwards they don't like the content.
This is a basic incentive problem that we can address by penalizing borderline content so it gets less distribution and engagement. By making the distribution curve look like the graph below where distribution declines as content gets more sensational, people are disincentivized from creating provocative content that is as close to the line as possible.

This process for adjusting this curve is similar to what I described above for proactively identifying harmful content, but is now focused on identifying borderline content instead. We train AI systems to detect borderline content so we can distribute that content less.
The category we're most focused on is click-bait and misinformation. People consistently tell us these types of content make our services worse -- even though they engage with them. As I mentioned above, the most effective way to stop the spread of misinformation is to remove the fake accounts that generate it. The next most effective strategy is reducing its distribution and virality. (I wrote about these approaches in more detail in my note on Preparing for Elections.)
Interestingly, our research has found that this natural pattern of borderline content getting more engagement applies not only to news but to almost every category of content. For example, photos close to the line of nudity, like with revealing clothing or sexually suggestive positions, got more engagement on average before we changed the distribution curve to discourage this. The same goes for posts that don't come within our definition of hate speech but are still offensive. This pattern may apply to the groups people join and pages they follow as well. This is especially important to address because while social networks in general expose people to more diverse views, and while groups in general encourage inclusion and acceptance, divisive groups and pages can still fuel polarization. To manage this, we need to apply these distribution changes not only to feed ranking but to all of our recommendation systems for things you should join.
One common reaction is that rather than reducing distribution, we should simply move the line defining what is acceptable. In some cases this is worth considering, but it's important to

remember that won't address the underlying incentive problem, which is often the bigger issue. This engagement pattern seems to exist no matter where we draw the lines, so we need to change this incentive and not just remove content.

I believe these efforts on the underlying incentives in our systems are some of the most important work we're doing across the company. We've made significant progress in the last year, but we still have a lot of work ahead.

By fixing this incentive problem in our services, we believe it'll create a virtuous cycle: by reducing sensationalism of all forms, we'll create a healthier, less polarized discourse where more people feel safe participating.

Giving People Control and Allowing More Content

Once we have technology that can understand content well enough to proactively remove harmful content and reduce the distribution of borderline content, we can also use it to give people more control of what they see.

The first control we're building is about providing the safer experience described above. It will be on by default and it means you will see less content that is close to the line, even if it doesn't actually violate our standards. For those who want to make these decisions themselves, we believe they should have that choice since this content doesn't violate our standards.

Over time, these controls may also enable us to have more flexible standards in categories like nudity, where cultural norms are very different around the world and personal preferences vary. Of course, we're not going to offer controls to allow any content that could cause real world harm. And we won't be able to consider allowing more content until our artificial intelligence is accurate enough to remove it for everyone else who doesn't want to see it. So we will roll out further controls cautiously.

But by giving people individual control, we can better balance our principles of free expression and safety for everyone.

Addressing Algorithmic Bias

Everything we've discussed so far depends on building artificial intelligence systems that can proactively identify potentially harmful content so we can act on it more quickly. While I expect this technology to improve significantly, it will never be finished or perfect. With that in mind, I

will focus the rest of this note on governance and oversight, including how we handle mistakes, set policies, and most importantly increase transparency and independent review.

A fundamental question is how we can ensure that our systems are not biased in ways that treat people unfairly. There is an emerging academic field on algorithmic fairness at the intersection of ethics and artificial intelligence, and this year we started a major effort to work on these issues. Our goal is to develop a rigorous analytical framework and computational tools for ensuring that changes we make fit within a clear definition of fairness.

However, this is not simply an AI question because at a philosophical level, people do not broadly agree on how to define fairness. To demonstrate this, consider two common definitions: equality of treatment and equality of impact. Equality of treatment focuses on ensuring the rules are applied equally to everyone, whereas equality of impact focuses on ensuring the rules are defined and applied in a way that produces equal impact. It is often hard, if not impossible, to guarantee both. Focusing on equal treatment often produces disparate outcomes, and focusing on equal impact often requires disparate treatment. Either way a system could be accused of bias. This is not just a computational problem -- it's also an issue of ethics. Overall, this work is important and early, and we will update you as it progresses.

Building an Appeals Process

Any system that operates at scale will make errors, so how we handle those errors is important. This matters both for ensuring we're not mistakenly stifling people's voices or failing to keep people safe, and also for building a sense of legitimacy in the way we handle enforcement and community governance.

We began rolling out our content appeals process this year. We started by allowing you to appeal decisions that resulted in your content being taken down. Next we're working to expand this so you can appeal any decision on a report you filed as well. We're also working to provide more transparency into how policies were either violated or not.

In practice, one issue we've found is that content that was hard to judge correctly the first time is often also hard to judge correctly the second time as well. Still, this appeals process has already helped us correct a significant number of errors and we will continue to improve its accuracy over time.

Independent Governance and Oversight

As I've thought about these content issues, I've increasingly come to believe that Facebook should not make so many important decisions about free expression and safety on our own.

In the next year, we're planning to create a new way for people to appeal content decisions to an independent body, whose decisions would be transparent and binding. The purpose of this body would be to uphold the principle of giving people a voice while also recognizing the reality of keeping people safe.

I believe independence is important for a few reasons. First, it will prevent the concentration of too much decision-making within our teams. Second, it will create accountability and oversight. Third, it will provide assurance that these decisions are made in the best interests of our community and not for commercial reasons.

This is an incredibly important undertaking -- and we're still in the early stages of defining how this will work in practice. Starting today, we're beginning a consultation period to address the hardest questions, such as: how are members of the body selected? How do we ensure their independence from Facebook, but also their commitment to the principles they must uphold? How do people petition this body? How does the body pick which cases to hear from potentially millions of requests? As part of this consultation period, we will begin piloting these ideas in different regions of the world in the first half of 2019, with the aim of establishing this independent body by the end of the year.

Over time, I believe this body will play an important role in our overall governance. Just as our board of directors is accountable to our shareholders, this body would be focused only on our community. Both are important, and I believe will help us serve everyone better over the long term.

Creating Transparency and Enabling Research

Beyond formal oversight, a broader way to create accountability is to provide transparency into how our systems are performing so academics, journalists, and other experts can review our progress and help us improve. We are focused on two efforts: establishing quarterly transparency and enforcement reports and enabling more academic research.

In order to improve our systems, we've worked hard to measure how common harmful content is on our services and track our effectiveness over time. When we were starting to build and debug our measurement systems, we only used the data internally to focus our work. As we've gained confidence in the measurements of more of our systems, we're publishing these metrics as well so people can hold us accountable for our progress. We released our first transparency and

enforcement report earlier this year, and we're releasing the second report today, which you can read here.

These reports focus on three key questions:

1. How prevalent, or common, is content that violates our Community Standards? We think the most important measure of our effectiveness in managing a category of harmful content is how often a person encounters it. For example, we found that in the third quarter of this year between 0.23-0.27% of content viewed violates our policies against violent and graphic content. By focusing on prevalence, we're asserting that it's more important to remove a piece of harmful content that will be seen by many people than it is to quickly remove multiple pieces of content that won't be as widely viewed. We think prevalence should be the industry standard metric for measuring how platforms manage harmful content.

2. How much content do we take action on? While less important than prevalence, this still demonstrates the scale of the challenges we're dealing with. For example, in Q3 we removed more than 1.2 billion pieces of content for violating our spam policies. Even though we typically remove these before many people see them so prevalence is low, this shows the scale of the potential problem if our adversaries evolve faster than our defenses.

3. How much violating content do we find proactively before people report it? This is the clearest measure of our progress in proactively identifying harmful content. Ideally our systems would find all of it before people do, but for nuanced categories we think 90%+ is good. For example, 96% of the content we remove for nudity is identified by our systems before anyone reports it, and that number is 99% for terrorist content. Because these are adversarial systems, these metrics fluctuate depending on whether we're improving faster than people looking for weaknesses in our systems.

Our priority is getting these measurements stable enough to report for every category of harmful content. After that, we plan to add more metrics as well, including on mistakes we make and the speed of our actions.

By late next year, we expect to have our systems instrumented to release transparency and enforcement reports every quarter. I think it's important to report on these community issues at the same frequency as we report our earnings and business results -- since these issues matter just as much. To emphasize this equivalence further and to create more accountability, we will start doing conference calls just like our earnings calls after we issue each transparency report.

In addition to transparency reports, we're also working with members of the academic community in different ways to study our systems and their impact. This work already focuses on preventing misuse during elections as well as removing bad content from our services. We also plan to expand this work to share more information on our policy-making and appeals processes, as well

as working on additional research projects. These partnerships are critical for learning from outside experts on these important challenges.

Working Together on Regulation

While creating independent oversight and transparency is necessary, I believe the right regulations will also be an important part of a full system of content governance and enforcement. At the end of the day, services must respect local content laws, and I think everyone would benefit from greater clarity on how local governments expect content moderation to work in their countries.

I believe the ideal long term regulatory framework would focus on managing the prevalence of harmful content through proactive enforcement. This would mean defining the acceptable rates of different content types. Without clear definitions, people rely on individual examples of bad content to understand if a service is meeting its overall responsibilities. In reality, there will always be some harmful content, so it's important for society to agree on how to reduce that to a minimum -- and where the lines should be drawn between free expression and safety.

A good starting point would be to require internet companies to report the prevalence of harmful content on their services and then work to reduce that prevalence. Once all major services are reporting these metrics, we'll have a better sense as a society of what thresholds we should all work towards.

To start moving in this direction, we're working with several governments to establish these regulations. For example, as President Macron announced earlier this week, we are working with the French government on a new approach to content regulation. We'll also work with other governments as well, including hopefully with the European Commission to create a framework for Europe in the next couple of years.

Of course, there are clear risks to establishing regulations and many people have warned us against encouraging this. It would be a bad outcome if the regulations end up focusing on metrics other than prevalence that do not help to reduce harmful experiences, or if the regulations end up being overly prescriptive about how we must technically execute our content enforcement in a way that prevents us from doing our most effective work. It is also important that the regulations aren't so difficult to comply with that only incumbents are able to do so.

Despite these risks, I do not believe individual companies can or should be handling so many of these issues of free expression and public safety on their own. This will require working together across industry and governments to find the right balance and solutions together.

Conclusion

These questions of what we want the internet to be are some of the most important issues facing our society today. On one hand, giving people a voice aligns with our democratic ideals and enlightenment philosophy of free thought and free expression. We've seen many examples where giving people the power to share their experiences has supported important movements and brought people together. But on the other hand, we've also seen that some people will always seek to use this power to subvert these same ideals [to divide us]. We have seen that, left unchecked, they will attempt to interfere in elections, spread misinformation, and even incite violence.

There is no single solution to these challenges, and these are not problems you ever fully fix. But we can improve our systems over time, as we've shown over the last two years. We will continue making progress as we increase the effectiveness of our proactive enforcement and develop a more open, independent, and rigorous policy-making process. And we will continue working to ensure that our services are a positive force for bringing people closer together.