July 2022

Privacy within Meta's Integrity Systems

Why user rights are at the center of our safety and security approach

Meta

What does privacy mean when data a person creates or data observed about them can be used to protect a broader community?

This question is older than the Internet,[1] but feels particularly relevant for the modern day where many people spend a significant portion of their lives online. The Universal Declaration of Human Rights states that people have rights to both privacy and to safety, as well as numerous other rights like freedom of speech and association.[2] These values allow for autonomy, creation, and the many things that comprise living and growing in society. And while this charter was intended for and signed by governments, these values still offer important guideposts for what people desire, or even expect, out of their experiences with online companies.

But when it comes to social media, discourse around privacy, speech, and safety can sometimes pull these values into different camps or pit them against each other — when in reality each value is critical to enabling high-quality online interactions. A prime example of this is the ongoing discussion around decreasing hate speech and harassment online. Some experts have suggested that hate speech and harassment are amplified online because some social media platforms allow their users to choose their own names or be pseudonymous, and that social media companies instead should verify the identity of all their users.[3] Other experts and researchers challenge that suggestion, finding that non-anonymous individuals may actually be more aggressive online than anonymous individuals.[4] Further, global privacy laws have strict requirements for the collection of identity information that could conflict with some content moderation proposals.

For people who use Meta's family of apps — Facebook, Instagram, WhatsApp and Messenger — individual privacy protections must coexist alongside the voice and safety values of the community. People do not want to share their thoughts, photos, and details of their lives if they cannot control who can see and interact with that information. At the same time, many people do not feel comfortable joining a community or a conversation if they might be verbally attacked for their opinions or, even worse, because of biographic features like their gender or race. People rightly expect online services to prioritize both privacy and safety, but operating social services at scale in a way that appropriately serves both values is a difficult problem — and one that Meta cannot and should not try to solve on its own.

This paper is designed to bring more people together, into the same virtual space, to discuss how Meta can, and should, analyze privacy alongside the many different values that go into content moderation decisions. We want to solicit feedback on our approaches to privacy protections in the context of the challenging issues we face across Meta services about the safety of people on our platforms, the security of accounts, and the integrity of our platforms themselves. The case studies included here are not exhaustive of the work we do in this area, but they are some of the most frequent or most pressing problems we see on our services and illustrate how we believe we can often accomplish privacy and safety goals together in the same projects and tools. We hope this format creates an open dialogue to discuss what people want out of new and existing Meta services when it comes to privacy, speech, and safety.

Privacy is a core value in safety and security enforcement.

Across Meta, the teams that focus on integrity,[5] safety, and security design products with multiple user rights in mind, such as voice, privacy, safety, security, well-being, trust, authenticity, dignity, and fairness.[6] These different values all relate to creating positive experiences for people online where they feel free to connect and share. We believe that protecting people's privacy is key to creating positive experiences across Meta products and services because people want to know that their information is used and handled appropriately.

But how we approach privacy in safety and security product design is informed by the context of the problem at hand, and there are times when getting safety and security right means using information about a person to identify, remove, and prevent certain online experiences. For instance, on Facebook you might share details about yourself in your "about me" section on your profile, you might join many groups but only post in and participate in conversations in a few, and when you interact with Facebook we receive details about your IP address and device. These points of information may be necessary to address certain safety or security challenges, but may not be relevant for others. For example, an account sign-in from a new location or device can be helpful to spot if an account has been compromised, but it may have little relevance for determining if a picture is nudity or someone in a bathing suit. And your "about me" and group activity can be useful to detect accounts engaging in policy-violating activity, but may not be necessary to determine whether a post contains facts or harmful misinformation.[7]

Before diving into the case studies, we will explore some of the context that drives our thought process and approach to privacy within safety and security tools at Meta: 1) the safety, security, and integrity issues we see across Meta, 2) what people and governments are asking social media companies to do on both privacy and safety, and 3) the process where we assess privacy concerns and ensure adequate protections in tools built for safety.

Meta is committed to reducing bad experiences on our services.

The kind of harms and negative experiences that Meta seeks to prevent on our services through our Community Standards are not new, not unique to the internet, and not unique to Meta.[8] Academics, regulators, and non-profit organizations have been tackling questions of safety and free speech online for decades.[9] Data usage and data protections have not often been discussed in these conversations about how to moderate online dialogue, but with advances in technology like machine learning that can automatically detect and take down violating content, that is changing.[10]

The expansion of digital spaces in which we increasingly interact have created new opportunities for bad actors to exploit peoples' safety, security, and well-being online, ranging from fraud and online crimes to the promotion of violence, the spread of harmful misinformation, and fostering of hate.[11] For Meta and others that want to decrease these negative experiences online, we need to process content and sometimes other personal data in order to detect and take down violating content and accounts.

Across Meta, we've developed the Community Standards as well as platform-specific policies that detail how we expect individuals to act in order to create positive experiences for themselves and the entire community.[12] These policies have been developed over the last decade, with multiple rounds of consultation with external experts in areas like hate speech, child safety, and privacy.[13] Each of these policies have been informed by how online problems manifest on the platform, and they continue to evolve to meet new challenges. The policies list the circumstances in which we take action on content, accounts, or different functions within each service.

For example, harassment is prohibited on Facebook, Instagram, and Messenger through the Community Standards,[14] as well as Instagram's Community Guidelines.[15] The process on each platform looks slightly different depending on people's expectations, and on the data and tools available. For instance, Facebook and Instagram may remove a post containing harassment after using automation to detect language associated with bullying, and Facebook might also reduce the distribution of borderline harassment content and content where we suspect, but have not confirmed, it may be harassing.[16] On Messenger, a person can choose to report data from their device, including sending Meta the most recent messages sent in that conversation. This allows us to take action if violations are detected in messages.[17]

The regulatory environment for privacy, free speech, and safety is shifting.

Over the past few decades, the world has experienced an exponential increase in the amount of data generated, processed and stored about its global citizens.[18] Alongside these social and economic changes, it has become all the more important to recognize and protect individual privacy rights. An ever-growing set of global privacy regulations (including the European Union's General Data Protection Regulation ("GDPR"),[19] Brazil's Lei Geral de Proteção de Dados ("LGPD"),[20] and Japan's Act on the Protection of Personal Information[21]) have emerged to govern the way companies collect, process, store and transfer personal data. The meaning of "personal data" likewise has expanded far beyond traditional identifiers like a person's name and contact information and includes the content people post online as well as information about their online activity.

For the safety, security, and integrity space in particular, a concept that has taken root since the passage of the GDPR and LGPD in Brazil,[22] is a balancing test that requires technology companies to analyze whether the use of personal data is necessary and proportionate when it is used to prevent bad experiences and improve well-being on social media platforms.[23] This balancing act puts an obligation on companies like Meta to justify the use of data, even to achieve public online safety goals.

At the same time, governments are increasingly concerned about the negative experiences their citizens may face online. Newly proposed laws in India,[24] Australia,[25] and the United Kingdom[26] create new processes for governments to request that technology companies remove pieces of content. The European Union's terrorist content regulation goes a step further and explicitly encourages platforms to invest in automated content moderation to decrease the dissemination of terrorist content online.[27] And India's new Information Technology Rules, 2021 even creates a requirement for social media to publicly indicate we have "verified" all users who voluntarily share identity information, like a phone number, with Meta.[28]

These trends in global privacy and technology regulation create a need to evaluate use of personal data in projects individually not only to ensure compliance with safety regulations but also to ensure the use of data is justified under the necessary and proportionate standard.

Meta's Privacy Review offers a process to analyze privacy alongside other safety, security, and integrity concerns.

Across Meta, new products that touch user data, including internal tools, go through privacy review, where internal privacy experts across legal, policy, and product teams evaluate privacy risks associated with the project and determine if there are any changes that need to happen before launch to control for those risks.[29]

This review is especially important for safety, security and integrity tools where we might use a range of data to detect and prevent harm, such as the content you post, your communications with others and other information you provide when you use Meta products, including when you sign up for an account.[30] The data we use for integrity purposes can include metadata, such as the location of a photo or the date a file was created; information about your device or network connection, such as your device's operating system or your IP address or connection speed; information about the people, accounts, hashtags and Facebook groups, and Pages you are connected to and how you interact with them across our products; and information that we get from partners.[31] Through privacy review, a cross-functional team of experts can make decisions about what types of data are most appropriate to use in different safety, security, and integrity scenarios.

When teams discuss appropriate use of personal data, they analyze whether a project has built in protections across 8 different privacy principles:[32]

1. Purpose Limitation: Process data only for a limited, clearly stated purpose that provides value to people.
2. Data Minimization: Collect and create the minimum amount of data required to support clearly stated purposes.
3. Data Retention: Keep data for only as long as it is actually required to support clearly stated purposes.
4. External Data Misuse: Protect data from abuse, accidental loss and access by unauthorized third parties.
5. Transparency and Control: Communicate product behavior and data practices proactively, clearly and honestly. Whenever possible and appropriate, give people control over our practices.

6. Data Access and Management: Provide people the ability to access and manage the data that we have collected or created about them.
7. Fairness: Build products that identify and mitigate risk for vulnerable populations, and ensure value is created for people.
8. Accountability: Maintain internal process and technical controls across our decisions, products and practices.

In our evaluation of safety, security, and integrity tools, we also ask 1) whether the data is necessary to solve or understand the stated problem and 2) whether the data used is proportionate, taking into consideration the privacy expectations of our users and the severity of the problem we are trying to solve. Understanding necessity and proportionality for each integrity project helps us better evaluate Meta's privacy principles. Additionally, through our fairness principle, we evaluate whether projects and the privacy protections we build are inclusive of our global community.

By taking this case-by-case approach to evaluate privacy risks across Meta, our use of personal data is informed by and dependent on context, including which policies are violated, people's expectations on how Meta should stop those violations, what personal data is used in the project plan, and people's sentiments on how sensitive or private that data is. The case studies below should give insight into the different privacy and overall policy considerations we make for different problems we might see on Meta's platforms. These case studies certainly are not exhaustive of the range of negative experiences and harm we try to prevent, but are distinct and should highlight the balancing of interests that goes into each evaluation.

Case study: reducing hate speech while protecting privacy of personal data

Hate speech is prohibited on Meta's family of apps under our Community Standards,[33] but the problem of online hate speech is often in the news and top-of-mind for global policymakers as they are thinking about the dignity and well-being of their citizens online. While the vast majority of content on Facebook and Instagram is benign, if you are on the receiving end of hate speech, it can feel pervasive and naturally impacts people's desire to share their voice and engage. The negative impact hate speech has on individuals, communities, and our society is why we want to quickly detect and remove it through automation as soon as it is posted before many people can see it, rather than waiting for user reports after it has gotten many views.

Thinking about hate speech from a data perspective, hate speech is primarily content-based. Although the problem feels very personal between people, it is often reflected in words like racial slurs or images like nooses or swastikas that can be picked out of a post or interaction and identified as hate speech. As a result, focusing detection of violations against individual pieces of content, rather than detection of people, is generally the most effective way to address this challenge. If we tried to identify or predict people who might engage in hate speech — for example by building a model to predict the type of person who might post hate speech — we would run into accuracy and fairness issues.[34] It would likely exhibit unacceptable biases and not be able to account for the nuance of discrimination and hate in every country around the world. This is because there is no singular profile of a hate speaker or recipient. Hate speech varies from country to country or even group to group.[35] However, we have found success in predicting content that is likely hate speech by comparing previously-confirmed hate speech to new instances we see on our platforms.

Data Minimization

When we look at hate speech detection and enforcement through a privacy lens, we want to minimize data used in our automated enforcement to the most relevant data, and in this case, text and images in previously-confirmed instances of hate speech are often the most relevant to determining if a new piece of content is hate speech. At Meta, we generate this data in multiple ways: human reviewers label violating content as hate speech, subject matter experts create lists of hate speech phrases for specific languages and countries, and after these initial inputs, we can build automated tools to detect and remove hate speech on the platform.[36] The content information that our machine learning models for hate speech are primarily trained on can include things like looking at the language or where in the world it is getting views, but generally we can assess this information by reviewing the content itself without inquiring into the person who posted the content.

Data Retention

In addition to minimizing personal data used in automated enforcement, privacy review looks at privacy protective storage options and aims to limit the data stored to only that which is necessary to continue improving our ability to identify hate speech on our services. For hate speech, we may store violating content we have removed for two years for model training. This allows us to detect reposts of previously determined hate speech, which we have found can come up cyclically around major events like police shootings or elections. What can be less relevant to future detection is precise details of who originally posted past violating content like their name or other identifiers. While we may initially collect this information in order to take down the violating hate speech and assign penalties,[37] personal identifiers are typically not used to train automated hate speech detection.

Fairness

Another privacy concern when it comes to automated enforcement of hate speech is building procedural fairness into content removal and other actions we take. We support people's freedom of expression and therefore believe that it is most fair to remove hate speech content when there is a very high likelihood that it is violating, and at a likelihood where our measurement systems find few false positives.[38] To increase the accuracy of our decisions, the models that we've built from past instances of hate speech essentially look for similarities with new content that comes through the platform and then score the likelihood that the new content is hate speech. We strive to make this automated removal as accurate, if not more accurate, than human removal.

A challenge we have long faced, however, is that some content has indications that it violates our community standards but where we are not confident enough to remove it automatically. On Facebook, we've begun lowering the visibility of this type of content on the platform until it can be reviewed by a human moderator, which we believe creates a better environment for the community at large while minimizing the impact to individuals and their voice.[39]

Discussion Questions

1. Do you agree with our assessment that hate speech is primarily a content-based problem, or should we explore more tactics to detect and target the people who post hate speech?
2. What factors should Meta consider for a global population when evaluating our hate speech enforcement for algorithmic fairness? Would it be proportionate to use data about people who post or see hate speech in order to understand the fairness of our hate speech models across different communities?

Case study: reducing nudity and preserving personal control over images

Most adult nudity and sexual activity are also prohibited on Facebook and Instagram.[40] A challenge for this space is that sometimes adult nudity and sexual activity can reflect more severe violations like non-consensual intimate imagery, but it can be difficult to differentiate between these cases, both from an automation standpoint and even in human review if we have limited information. As a result, we needed to build additional protections to help people address cases where they have intimate images of them shared without consent.

Fairness

It can be incredibly difficult, without direct information from the subject of an image or video, to know whether or not adult nudity and sexual activity was consensually taken and consensually shared on our platforms.

One way we can account for this lack of knowledge, however, is by using automation to remove all forms of adult nudity that violate Community Standards on Instagram and Facebook. In this case, automation can potentially provide a more privacy-preserving review for people in vulnerable situations because we can proactively find and remove this content before it is viewed and reported by others, ultimately resulting in less eyes on it.[41]

The technology we use to detect intimate photos (image processing and media match software) also respects people's desire for intimate photos to generally stay private. The algorithms look for indications of nudity or sexual activity in images and videos because they have been trained on previous violations of Meta policies. When we are confident in a match and able to remove adult nudity automatically, the content does not go to human review, which also results in is fewer human eyes on potentially nude or sexual pieces of media.

External Data Misuse

Because our automated removal of adult nudity and sexual activity is trained on known violations of the Meta Community Standards, one thing people can do to prevent the spread of their non-consensual intimate imagery is help train our models to spot their images and videos.

Meta has partnered with StopNCII.org to offer a solution that empowers people around the globe to proactively thwart the non-consensual sharing of their intimate images.[42] People can create a case with StopNCII.org, which assigns a unique hash value (a numerical code) to their image, creating a secure digital fingerprint. Tech companies participating in StopNCII.org receive the hash and can use that hash to detect if someone has shared or is trying to share those images on their platforms. But the original image never leaves the person's device. Only hashes, not the images themselves, are shared with StopNCII.org and participating tech platforms.[43] This feature prevents further circulation of that NCII content and keeps those images securely in the possession of the owner.

Discussion Questions

1. Can high-precision automation be a privacy tool when it results in fewer human eyes on content that people generally view as sensitive or intimate?
2. Are there other negative experiences online where the StopNCII.org model (hash matching facilitated through an NGO) would create a better experience for victims?

Case study: retaining disabled account data to increase child safety

An ongoing challenge for online platforms and services is preventing people who have already been given penalties for violating a platform's policies from starting a brand new account, particularly in the context of child safety.

We have legal obligations to report child sexual exploitation as defined under U.S. federal law and we also have broader policies under our Community Standards that prohibit child sexual abuse materials, content that sexualizes minors, and inappropriate interactions with children.[44] In order to prevent any users whom we have disabled for these violations from coming back to our platforms with new accounts, we must retain certain account data to prevent new account sign-ups and to detect account holders who are repeatedly violating our policies.

Necessity and Proportionality

Child sexual exploitation violations are one of the most severe policy violations that a user can commit, and we believe that using relevant personal account data in addition to content is proportionate when protecting children by reducing recidivism. First, when we detect severe child exploitative imagery, we may disable the account that shared it and then, as required by law, report both the content as well as the account holder to the National Center for Missing and Exploited Children (NCMEC) who forwards those reports to law enforcement agencies around the world. We retain the report and relevant account information for a period of time, as required by law.[45] We also may retain some account and device information, like IP addresses, of accounts that violate our child safety policies in order to promote safety, security, and integrity, including using this information to prevent the creation of new accounts.[46]

Our decision to retain this information is rooted in the belief that violating our child safety policies is so severe that we are justified in using account information from child safety-disabled accounts to proactively prevent that person from continuing their behavior through a new account. We don't typically retain this type of information from deleted, non-violating accounts — for instance, when you delete your Facebook account, we permanently delete your profile, photos, and everything you have added to the account.[47] Ultimately, when it comes to child safety we offer less privacy protections to known violators of our policies, but we believe that retention and use of personal data in this context is proportionate because of the benefit gained by preventing these violators from accessing our services.

Discussion Questions

1. Are there other problem areas where it might be proportionate to retain device info in order to prevent new accounts of known policy violators?
2. How should proportionality and severity of harm intersect? What considerations should factor into problem-by-problem analysis?

Case study: collecting identity information to protect integrity of accounts

Identity information is viewed as sensitive and private by many people around the world. But there are a number of instances where it is necessary from a safety and account security perspective for Meta to ask who is behind an account.[48] To balance these competing values, we limit when we request identity information and how we store it.

When we believe that an account has been compromised or hacked on Facebook and Instagram, we ask people to verify their identity before they can keep using the account. This verification is necessary to make sure we only give access to the correct owner.[49] People may be able to retrieve their account with their phone number to get a security code or people may be able to regain access by sending in their government identification document to Meta.

Transparency and Control

We recognize that using a person's phone number or government ID to restore access to their account may cause people to worry about their privacy. Because of this, we believe it is important to give people control of this information after they submit it. People can go into their Account Center to change or delete the security phone number that they share with Facebook. And people can view IDs they have submitted during a security check in Account Center to control retention. By default, IDs are typically stored for 1 year on Facebook.[50] This allows security teams to validate account owners and train models that look for fakes and other potential problems with IDs. Anyone who has submitted an ID to Facebook can opt-out of training and request that their ID only be stored for 30 days in order to verify their account.

It is important to note that while authentic identity is a Community Standard and required on Facebook,[51] an exact match of a profile or account to a government identity isn't the only definition of authenticity. Online, people have more freedom to emphasize different aspects of who they are. For example, a person may want to create a Page for their business, and connect with different audiences. We also recognize that many people may not go by the name on their birth certificate or government document for a number of reasons, including if a person is transgender. And globally, nearly 1 billion people do not have access to government ID documents.[52] We're exploring new ways of verifying identity that aren't dependent on government ID,[53] but given current limitations, and the concerns people may reasonably have about requests to share identity information, it is all the more important for Meta requests for government ID to be narrow and tailored.

Discussion Questions

1. Does the collection of identity information in limited circumstances make sense?
2. As Meta explores other identity solutions, what considerations should we take into account in addition to privacy and security?

Case study: increasing safety in times of crisis

One of the most challenging issues for the teams that work on safety, security, and integrity across Meta is how to respond in the moment to global events and times of crisis. In these times we have at our disposal different tools to analyze and address what is happening, and we may take different actions depending on events as they unfold. In the last year, we've developed new tactics to protect people's privacy and security in response to military actions in Afghanistan and Ukraine, which this case study will discuss.

First, Meta has prioritized hiring global safety and human rights experts and works closely with on-the-ground partners and nonprofits to understand the unique considerations for different civic events for different countries around the world. We have less ability to offer meaningful services if we don't have information about what is happening to people in the region.

When we adjust our existing tools and design new tools for crisis events, we try to balance the need for people to express themselves with the fact that violent rhetoric may be increasing in the region as well as with the fact that people may also be targeted by others for their expression. What this means is that we may pull time-limited, emergency levers that reduce and remove content on the platform or affect people's ability to connect because we believe that the risk of harm is so great to justify limiting more speech on the platform than we otherwise would.

Alternatively, we also might adjust our policies to allow for emerging speech, such as Ukrainian expression of self-defense and calls for violence against Russian invading forces.[54]

Necessity and Proportionality

As events unfolded in both Afghanistan and Ukraine, we realized that activists, journalists, and people speaking out about invading military forces in the countries faced significant risk.[55] One option we had to help increase protection for these people was to limit their public visibility on both Facebook and Instagram. But to do this, we needed to use people's location information to tailor and promote privacy and security tools to them. We used people's stated location, IP address, or precise location if they had location services turned on,[56] but often, our use of location data is up-leveled to country or region because this allows us to take safety actions while preserving the privacy (and safety) of precise location.

Transparency and Control

In developing features to protect people in Afghanistan and Ukraine, it was important to us that people maintain control over their account settings where it makes most sense. So on Facebook, we promoted the ability to "lock" your Facebook profile, a one-click tool to people in Afghanistan that can lock down their Facebook account so that non-friends could not download their profile picture or see any posts on their timeline.[57] And on Instagram, we promoted the ability to update your settings to make your profile private.[58] Generally, we believe it is a better experience for people to make their own decisions about how public their profile is, but one area where we did decide to remove visibility was removing public friend's lists in Afghanistan and Ukraine because people do not have control over choices that their friends make.[59]

For the war in Ukraine, we had reports that many people were active on Instagram DMs so we also decided to move up new user controls for messaging privacy and security. We rolled out one-to-one encrypted chats in Instagram direct messaging for adults in Russia and Ukraine,[60] and notified users that they have encrypted chat options. We also wanted to be mindful of the fact that adversarial users can use chat functions to compromise people's security and gain personal information about them and that Meta may not be able to see and intervene in this scenario if people are using encrypted messaging, so we also sent out alerts on Instagram when people get new message requests to raise awareness about impersonators and bots.[61]

Overall, these privacy-to-increase-safety features change as we at Meta learn more information about situations on the ground and how people in these countries going through crisis use our products. As we discuss new privacy or security tools, we weigh if the tool will impact a person's ability to connect with others, we weigh how we can give people the most control over whether or not they use the tool, and we discuss how to give as much education and transparency for these new tools as possible in the product experience. Overtime we may decide to turn some of the levers that we pull during crisis scenarios into global product experiences, but we also may decide that some levers should be limited to worst case scenarios. This is an experience that we will continue to refine with experts inside and outside of Meta so that we can be responsive to what people need the most.

Discussion Questions

1. Should technology companies adjust their policy enforcement approach to respond to crisis events?
2. What principles should Meta and other technology companies use to determine if an event or circumstance rises to the level of a crisis?

Evolving integrity: new obstacles for the Metaverse

Each year, new advances in technology and changes in our world bring new questions and challenges for what safety, security and integrity means for Meta's products and services and the people who use them. Facebook, Instagram, Messenger and WhatsApp continue to evolve with people's privacy, safety, and functionality expectations for these services as well as with new legislation and changes in the market.

We believe that putting people at the center of our Integrity approach results in a better experience for all people in our community. This means that for each new challenge, we need to evolve and be at the forefront of change in order to understand the needs and desires of people who use our products.

We’re at the very start of our journey of building toward the metaverse. Many of the products will only be fully realized in 5-10 years. But right now, we are planting the seeds for new, but continuous, conversations around how privacy, safety, voice, and other human rights should all intersect in not only a virtual world, but in a diverse and decentralized online experience where many companies provide services.

As we build these technologies, we know we have an obligation to take a responsible approach ourselves, explain how data is used, and enable people to control the things that matter most. Our approach will focus on minimizing the amount of data we collect, leveraging privacy-enhancing techniques like differential privacy or on-device processing, and building meaningful transparency and control into our products. We'll also need platform-level tools to help people manage their experiences and take action if they see something they're not comfortable with. Providing these tools in our first party experiences — and enabling other companies to build their own approaches — will be critical.

We’re also cognizant that many of the safety and security issues we see on Facebook and Instagram may look wholly different in the Metaverse. For instance, animal sales are prohibited on Facebook and Instagram because we want to avoid real-world harm to animals through things like puppy mills or exotic animal sales, but it may make sense for creators to develop virtual animals or virtual pets that can be sold in VR where this real-world problem may not exist. Safety is one of many areas where we will want to engage with experts to establish rules of the road for the first party experiences that Meta may provide in augmented reality and virtual reality.

Meta's Ask

We hope that this paper sets the table and facilitates conversations among privacy and content moderation experts, technology companies, and regulatory bodies on what we as a global community expect privacy to look like in the face of online safety, security, and integrity challenges. We expect these challenges to only grow, and scale with emerging technologies, including advances in artificial intelligence and virtual reality.

Meta isn't the only company facing these challenges, and we see this paper as an opportunity to learn from industry peers and receive feedback from global experts on free expression, human rights, public safety, online speech harms, and privacy. We will continue to engage on this subject matter, but if you have feedback on the paper itself or important questions you feel should be addressed, reach out to privacy4integrity@fb.com through September 30, 2022.

We welcome comments on this topic, and will use the information we receive to inform the ways in which we use personal information to detect and remove Community Standards violations.

Endnotes

1. Solove, Daniel J., A Brief History of Information Privacy Law. PROSKAUER ON PRIVACY, PLI, 2016, GWU Law School Public Law Research Paper No. 215, Available at SSRN: https://ssrn.com/abstract=914271.

2. Article 12 of the Universal Declaration of Human Rights requires that: "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks." United Nations, Booklet, Universal Declaration of Human Rights, https://www.un.org/en/udhrbook/pdf/udhr_booklet_en_web.pdf.

3. Nazia Parveen and David Tindall, "Ministers promise crackdown on online racism against black footballers," The Guardian (Jan. 31, 2021), https://www.theguardian.com/world/2021/jan/31/online-racism-black-footballers-marcusrashford-social-media; "Government considering 100 points of ID to get a Facebook, Tinder account," news.com.au (April 2, 2021), https://www.news.com.au/technology/online/security/government-considering-100-points-of-id-to-get-facebook-tinder-account/news-story/624550c621d662da7d3bd98ff3f0e888; Tim Cushing, "Australia Government Considers Stripping Internet Users of their Anonymity," Techdirt (April 6, 2021), https://www.techdirt.com/articles/20210405/23025046557/australian-government-proposes-stripping-internet-users-their-anonymity.shtml.

4. "in the context of online firestorms, non-anonymous individuals are more aggressive compared to anonymous individuals" Kajita Rost and Lea Stahel, "Digital Social Norms Enforcement: Online Firestorms in Social Media," (June 17, 2016), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0155923; "women [who face more harassment online] are much more likely to adopt temporary identities than men," Alex Leavitt, "This is a Throwaway Account: Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community," In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pages 317-327. (ACM 2015), https://dl.acm.org/doi/10.1145/2675133.2675175; "99% of the accounts suspended were not anonymous." Twitter, "Combatting Online Racist Abuse: An update following the Euros," (Aug 10, 2021), https://blog.twitter.com/en_gb/topics/company/2020/combatting-online-racist-abuse-an-update-following-the-euros.

5. Integrity is a term of art inside Meta. Trust and safety, computer and account security, the reduction of bad experiences, related privacy issues, and more are handled by a network of teams within Meta that often take “Integrity” as part of their name. Among many things, these teams build tools to prevent harm, moderate our platforms, and enforce our policies.
6. Meta’s Community Standards highlight five core principles (voice, authenticity, safety, privacy, dignity) for how we moderate the platform, but when you dive into the text, all of these rights are present. https://transparency.fb.com/en-gb/policies/community-standards/.
7. Meta removes harmful misinformation under the Community Standards. This includes misinformation that may lead to physical harm or violence, misinformation about vaccines and health during public health emergencies, misinformation that contributes to interference with people’s participation in political processes, and manipulated media that is intended to deceive people. https://transparency.fb.com/policies/community-standards/misinformation/.
8. Id. The Community Standards detail prohibited content, actions, and behavior like violence and criminal behavior, safety violations, objectionable content, integrity and authenticity violations, intellectual property violations, and removal of underage accounts.
9. See e.g., Reno v. American Civil Liberties Union, 521 U.S. 844 (1997); The Santa Clara Principles, https://santaclaraprinciples.org/; Danielle Keats Citron and Helen Norton, Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age, 91 B.U. L. REV. 1435 (2011), https://scholar.law.colorado.edu/cgi/viewcontent.cgi?article=1179&context=articles; Mary Anne Franks & Ari Ezra Waldman, “Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions,” 78 Md. L. Rev. 892 (2019), https://digitalcommons.law.umaryland.edu/cgi/viewcontent.cgi?article=3835&context=mlr;
10. See Eva Maydell, Member of European Parliament, remarks, “Leveraging AI: Risks & Innovation in Content Moderation by Social Media Platforms,” Computers, Privacy & Data Protection 2022, https://www.cpdpconferences.org/cpdp-panels/leveraging-ai-risks-innovation-in-content-moderation-by-social-media-platforms; U.S. Federal Trade Commission, Report to Congress, “Combatting Online Harms Through Innovation,” (June 16, 2022), https://www.ftc.gov/system/files/ftc_gov/pdf/Combatting%20Online%20Harms%20Through%20Innovation%3B%20Federal%20Trade%20Commission%20Report%20to%20Congress.pdf.

11. The majority of content posted by people on Facebook and Instagram does not violate the Community Standards or other policies, and we are often able to automatically remove many pieces of violating content before they are seen and reported by people on the platform. For example, in our recent Community Standards Enforcement Report, we removed 17.4 million pieces of hate speech content from October to December 2021, reducing the prevalence of hate speech to 2-3 pieces on the platform for every 10,000 pieces of content. But our human review and automated efforts are not infallible. Q4 2021 Report, https://transparency.fb.com/data/community-standards-enforcement/. "It's important to note that the vast majority of what people see on Facebook is neither political nor hateful. Political posts make up only about 6 percent of what people in the United States see in their News Feed, and the prevalence of hateful content people see on our service is less than 0.08 percent. While we work hard to prevent abuse of our platform, conversations online will always reflect the conversations taking place in living rooms, on television, and in text messages and phone calls across the country. Our society is deeply divided, and we see that on our services too." Mark Zuckerberg Testimony U.S. House of Representatives, House Energy and Commerce (March 25, 2021) Page 8, https://energycommerce.house.gov/sites/democrats.energycommerce.house.gov/files/documents/Witness%20Testimony_Zuckerberg_CAT_CPC_2021.03.25.pdf

12. See Community Standards at https://transparency.fb.com/policies/community-standards/; Other Policies, Meta Transparency Center, https://transparency.fb.com/policies/other-policies.

13. "Publishing our Internal Enforcement Guidelines and Expanding our Appeal Process," Meta Newsroom (April 24, 2018), https://about.fb.com/news/2018/04/comprehensive-community-standards/.

14. Community Standards, "Bullying and Harassment," https://transparency.fb.com/policies/community-standards/bullying-harassment/.

15. Instagram, Community Guidelines, https://www.facebook.com/help/instagram/477434105621119.

16. "Since 2016, the strategy of Integrity has been "remove, reduce, and inform" to manage problematic content across the Facebook family of apps. This involves removing content that violates our policies, reducing the spread of problematic content that does not violate our policies and informing people with additional information so they can choose what to click, read or share. This strategy applies not only during critical times like elections, but year-round." Facebook Newsroom, "Remove, Reduce, Inform: New Steps to Remove Problematic Content" (April 10, 2019), https://about.fb.com/news/2019/04/remove-reduce-inform-new-steps/.

17. Meta, “Meta’s Approach to Safer Private Messaging on Messenger and Instagram Direct Messaging,” (April 2022),
https://messengernews.fb.com/wp-content/uploads/2021/12/Metas-approach-to-safer-private-messaging-on-MSGR-and-IG-DMs-4.pdf.
18. Obama Administration, Big Data: Seizing Opportunities, Preserving Values (May 2014),
https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.
19. European Union, General Data Protection Regulation,
https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.
20. IAPP, Translation, Brazilian General Data Protection Law, As amended by Law No. 13,853/2019,
https://iapp.org/media/pdf/resource_center/Brazilian_General_Data_Protection_Law.pdf.
21. Japan Personal Information Protection Commission, Translation, Amended Act on the Protection of Personal Information, (Approved June 5, 2020)
https://www.ppc.go.jp/files/pdf/APPI_english.pdf.
22. Canada’s proposed Consumer Privacy Protection Act would also incorporate necessary and proportionate principles into companies’ obligations for the use of data in public interest scenarios. Speech by Daniel Therrien, Privacy Commissioner of Canada, “The Future of Privacy Law Reform in Canada,” Remarks at IAPP Canada Privacy Symposium 2021 (May 26, 2021),
https://www.priv.gc.ca/en/opc-news/speeches/2021/sp-d_20210526/.
23. European Data Protection Supervisor, “Necessity and Proportionality,”
https://edps.europa.eu/data-protection/our-work/subjects/necessity-proportionality_en.
Coalition of NGOs, “Necessary & Proportionate: On the Application of Human Rights to Communications Surveillance,” (July 10, 2013),
https://necessaryandproportionate.org/principles/.
24. Sheikh Saalik and Krutika Pathi, “India Internet Law adds fears over free speech, privacy,” AP News (July 14, 2021),
https://apnews.com/article/technology-entertainment-business-music-india-2458a729cff255c8a8f83d84101372d8.
25. Zachary Forrai, “A new era for Australian online safety regulation,” JD Supra (Aug 2, 2021),
https://www.jdsupra.com/legalnews/a-new-era-for-australian-online-safety-4740569/.
26. “Online Safety Bills: New Offenses and Tighter Rules,” BBC News (Dec. 14, 2021),
https://www.bbc.com/news/technology-59638569.
27. European Union, Regulation (EU) 2021/784, “Addressing the dissemination of terrorist content online,” (effective June 7, 2022) https://eur-lex.europa.eu/eli/reg/2021/784/oj.

28. Tanu Banerjee, et. al., "India's new rules for Facebook, WhatsApp and other social media platforms explained in five points," Business Insider (Mar. 2, 2021), https://www.businessinsider.in/policy/news/indias-new-rules-for-facebook-whatsapp-and-other-social-media-platforms-explained-in-five-points/articleshow/81285658.cms.
29. Meta, "Privacy Progress Update," https://about.facebook.com/privacy-progress.
30. Meta Data Policy, https://www.facebook.com/policy (last updated Jan. 4, 2022) (new Meta Privacy Policy at https://www.facebook.com/privacy/policy (effective July 26, 2022)).
31. Id.
32. See Meta Privacy Progress Update at "02. Accountability in Practice: Privacy Review."
33. The Community Standards define Meta's working definition of hate speech, including detailing the different actions we may take for the different types of hate speech we see online. This policy sets out our framework for enforcement and is available for anyone to read. https://transparency.fb.com/policies/community-standards/hate-speech/.
34. It is important to note that when we remove violating content from Facebook, we often assign a penalty or a "strike" to the person who posted the content for violating the Community Standards. When people accumulate strikes they may face restrictions on their account. So we do not predict hate speakers, but we do have records of who has violated the hate speech policy in the past. https://transparency.fb.com/enforcement/taking-action/counting-strikes/.
35. "Training AI to Detect Hate Speech in the Real World," Meta AI (Nov. 19, 2020), https://ai.facebook.com/blog/training-ai-to-detect-hate-speech-in-the-real-world/.
36. We are continuously improving on our automated systems, seeking to use less data while also improving accuracy. Content is fundamental to train machine learning, but with automation advancements like Few-Shot-Learner, we can build better detection for new problems where there are less pieces of content to train traditional, more bespoke AI models. "Harmful content can evolve quickly. Our new AI system adapts to tackle it." Meta AI (Dec. 8, 2021), https://ai.facebook.com/blog/harmful-content-can-evolve-quickly-our-new-ai-system-adapts-to-tackle-it/.
37. Supra, note 34.
38. Typically, when we remove hate speech content, we also notify the person who posted the content about our decision and provide them with the ability to disagree with our decision.
39. Content Distribution Guidelines, https://transparency.fb.com/features/approach-to-ranking/types-of-content-we-demote/; Guy Rosen, "Hate Speech Prevalence has Dropped by Almost 50% on Facebook," Meta Newsroom (Oct. 17, 2021) https://about.fb.com/news/2021/10/hate-speech-prevalence-dropped-facebook/.

40. Adult Nudity and Sexual Activity Policy, Community Standards, https://transparency.fb.com/policies/community-standards/adult-nudity-sexual-activity/

41. From a user safety and well-being perspective, automated detection and removal of Community Standards violations makes enforcement of those violations more efficient and leads to better overall platform health. Between April and June 2021, Facebook actioned 32.8 million pieces of adult nudity and sexuality, 98.9% of it found through automation. It also helps us apply our Community Standards with more consistency, so that people from all walks of life are held to the same rules, making our enforcement of Community Standards (and use of personal data to enforce Community Standards) more fair.

42. Antigone Davis, "Strengthening our efforts against the spread of non-consensual intimate imagery," FB Newsroom (Dec. 2, 2021), https://about.fb.com/news/2021/12/strengthening-efforts-against-spread-of-non-consensual-intimate-images/#:~:text=StopNCII.org%20builds%20on%20technology,remove%20them%20after%20the%20fact.

43. Id.

44. Child Sexual Exploitation Policy, Community Standards, https://transparency.fb.com/policies/community-standards/child-sexual-exploitation-abuse-nudity/

45. See Meta Data Policy.

46. Id.

47. Help Center, "How do I permanently delete my Facebook account?" https://www.facebook.com/help/224562897555674.

48. Cybersecurity Policy, Community Standards, https://transparency.fb.com/policies/community-standards/cybersecurity/.

49. There are a number of other instances in which Meta may ask for identity information to authenticate an account, such as when we get a report that a person is underage. We also ask for identity information from some businesses or to verify the identity and location of people placing political ads. If people decline to provide identity information in this scenario, they can still access and use their Facebook account, but they might be restricted from placing ads. Facebook Business Help Center, "Get Authorized to Run Ads about Social Issues, Elections, or Politics," https://www.facebook.com/business/help/208949576550051?id=288762101909005.

50. Facebook Help Center, "What happens to your ID after you send it to Facebook?," https://www.facebook.com/help/155050237914643?helpref=faq_content.

51. Account Integrity and Authentic Identity, Community Standards, https://transparency.fb.com/policies/community-standards/account-integrity-and-authentic-identity/.

52. Vyjayanti T Desai, et. al., “The global identification challenge: Who are the 1 billion people without proof of identity?,” World Bank Blogs (April 25, 2018) https://blogs.worldbank.org/voices/global-identification-challenge-who-are-1-billion-people-without-proof-identity.

53. For Instagram accounts where we suspect that the owner may be under 18 years old, we have begun using Yoti, an age verification system that analyzes faces to determine age, and we have begun offering a social vouching option where users can request that friends over 18 years old verify their age. Meta Newsroom, “Introducing New Ways to Verify Age on Instagram,” (June 23, 2022) https://about.fb.com/news/2022/06/new-ways-to-verify-age-on-instagram/.

54. “Update March 11, 2022, President Global Affairs, Nick Clegg, statement,” “Meta’s Ongoing Efforts Regarding Russia’s Invasion of Ukraine,” Meta Newsroom, (Feb. 26, 2022), https://about.fb.com/news/2022/02/metas-ongoing-efforts-regarding-russias-invasion-of-ukraine.

55. Nathaniel Gleicher, Meta Head of Security Policy, Twitter Thread (Aug. 19, 2021), https://twitter.com/ngleicher/status/1428474000611573762?lang=en; “Meta’s Ongoing Efforts Regarding Russia’s Invasion of Ukraine,” Meta Newsroom, (Feb. 26, 2022)https://about.fb.com/news/2022/02/metas-ongoing-efforts-regarding-russias-invasion-of-ukraine/.

56. Facebook Data Policy, https://www.facebook.com/policy (last updated Jan. 4, 2022).

57. Help Center, “How do I lock my profile on Facebook?” https://www.facebook.com/help/196419427651178.

58. “Meta’s Ongoing Efforts Regarding Russia’s Invasion of Ukraine,” Meta Newsroom, (Feb. 26, 2022)https://about.fb.com/news/2022/02/metas-ongoing-efforts-regarding-russias-invasion-of-ukraine/ at “Originally Published Feb 26, 2022).

59. Supra note 44.

60. Id. at “Update February 28, 2022).

61. Id. at “Originally published February 26, 2022).