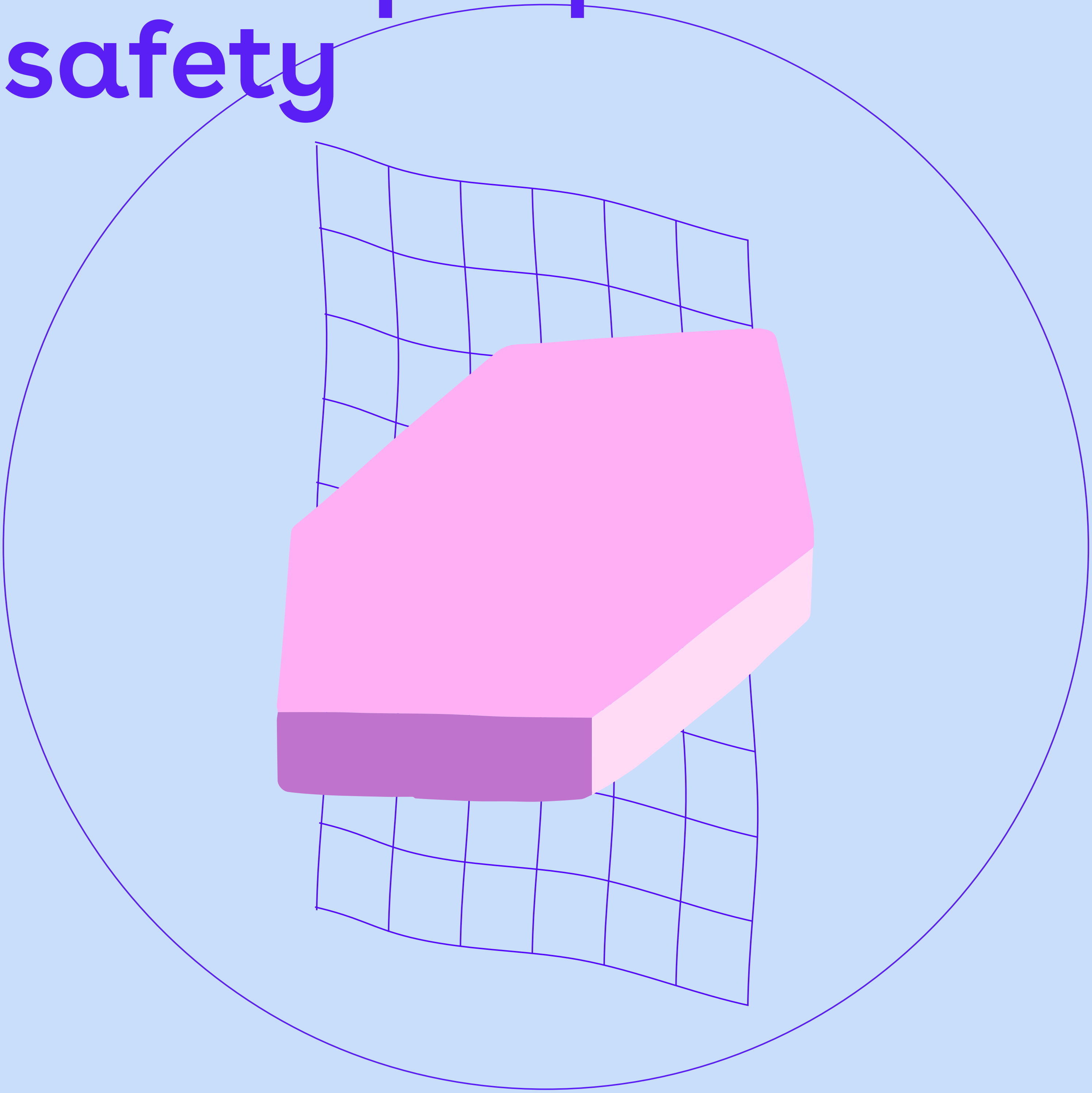
Welcome_

Civic Signals

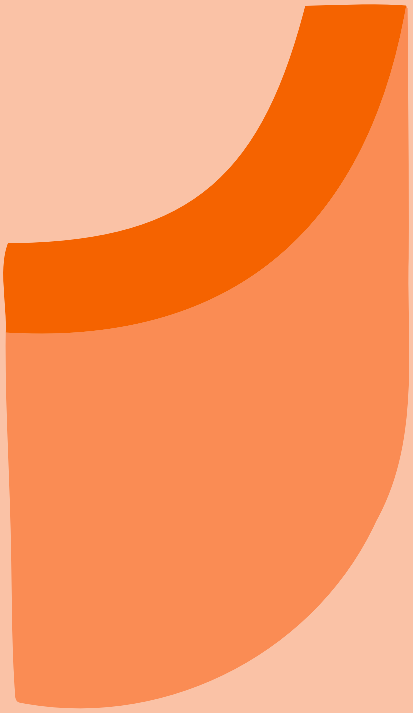
Ensure people's safety



This signal is part of Civic Signals, a larger framework to help create better digital public spaces. We believe it's a platform's responsibility to design the conditions that promote ideal digital public spaces. Such spaces should be designed to help people feel Welcome, to Connect, to Understand and to Act. These four categories encompass the 14 Civic Signals.

Table of contents

At a glance



Safety is a state of being protected from harm or danger, which can range from malware, identity theft, harassment and cyberbullying to internet addiction, sexual victimization, and exposure to violent material.

Why It Matters

When people feel safe, they get more out of their online interactions. Being safe means feeling protected from reputational and physical harms. In contrast, feeling unsafe can lead to withdrawal from online spaces. Online safety is particularly important to consider for vulnerable people such as children, and for marginalized groups, who can be further marginalized by threats to their safety.

> *Social networks might bring people together but can also pull them away, can bring evilness too, so we have to wise up… especially with kids."*
> *– Miguel, Brazilian focus group participant*

Putting the Signal Into Practice

- For both algorithmic and human content moderation, increased transparency in moderation decisions can help newcomers to learn a platform's rules, and lower the potential for violations, researcher Shagun Jhaver and collaborators have argued. https://www.cc.gatech.edu/~sjhaver3/Removal_Explanations.pdf

- Giving people a voice in rulemaking, and considering the community's social norms when determining rules, can result in more effective safety guidelines. For example, in addition to platform-wide guidelines, Reddit allows individual forums ("subreddits") to have their own rules tailored to their community around issues like harassment, hate speech, and inappropriate content. https://www.reddit.com/r/modnews/comments/42o2i0/moderators_subreddit_rules_now_available_for_all/

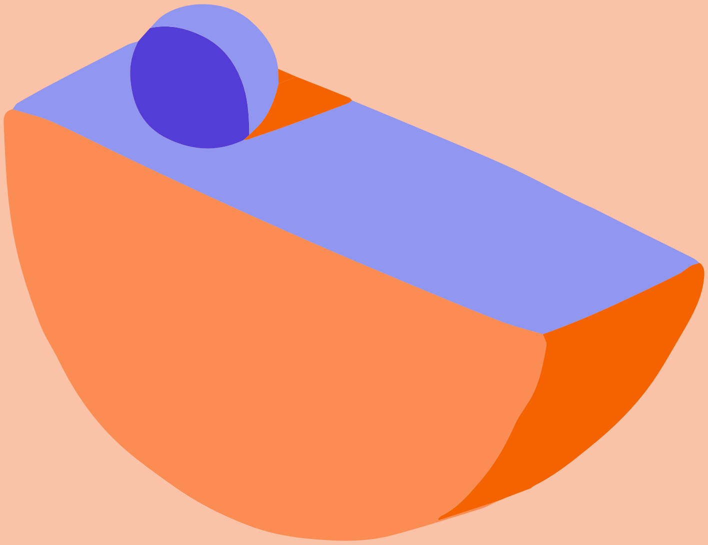
- Promoting safety doesn't always mean removing content. Tumblr displays an "Everything okay?" message of support, along with resources, when people search on the platform for content related to self-harm. https://support.tumblr.com/post/74751945752/everything-okay

- When it comes to parental control strategies, it's best if youth are involved through participatory design, work by researcher Brenna McNally and colleagues demonstrates. https://pearl.umd.edu/wp-content/uploads/2018/08/2018-McNally-CHI-ParentalMobileMonitoring-Paper.pdf

- Games and storytelling can be used to help children and adolescents understand online risks, as researcher Fotis Lazarinis and colleagues detail. Here are seven such games: https://www.makeuseof.com/tag/6-internet-safety-games-kids-cyber-smart/

- Common Sense links to a number of children's lesson plans about online safety. https://www.commonsense.org/education/articles/23-great-lesson-plans-for-internet-safety

Literature review



By Casey Fiesler,
University of Colorado Boulder

What the Signal Is

In its most basic definition, safety is a state of being protected from harm or danger. Harm, in turn, results from specific threats or vulnerabilities.

In an online context, discussions of online threats fall generally into a few major categories. Synthesized from various works, including those from social media scholar danah boyd and communication scholar Eszter Hargittai; informatics researcher Emmanouil Magkos and colleagues; and safety and security researcher Heidi Hartikainen and colleagues, these include: *contact threats* (physical harm, e.g., sexual victimization); *content threats* (psychological harm, e.g., exposure to violent material); *computer threats* (direct harm from tech; e.g., malware or internet addiction); and *commercial threats* (harm from manipulation, e.g., spam or phishing).

Safety in online spaces is best explained in the context of the types of harm that might occur; what do participants in that space need to be kept safe from?

For example, digital literacy education or monitoring tools for children and adolescents often focus on threats such as inappropriate content (e.g., pornography, graphic violence), predators, cyberbullying, phishing, screen addiction, and information security (e.g., password practices).

Adults in online spaces are at risk from many of these safety threats as well—for example, spam, phishing, hacking, identity theft, internet addiction, triggering content, online harassment, and privacy violations. Some harms are "felt" online only, and others occur offline, brought on by online threats—for example, doxxing (revealing someone's personal information online) leading to a real-world confrontation. Human-computer interaction (HCI) researchers Jessica Pater and Beth Mynatt have also pointed to the role that online communities play in the exposure and encouragement of self-harm.

Safety should also be considered not only at the individual but at the *collective* level, as emphasized by HCI researcher Morgan Klaus Scheuerman and colleagues in their analysis of online safety for the transgender community. Some harms have implications beyond the individual and instead affect entire communities. For example, online hate speech can contribute to the proliferation of racism, and disinformation is a threat to society as a whole. We discuss racial tolerance under our Encourage the Humanization of Others signal, and disinformation under Show Reliable Information.

Finally, it is important to note that participants in online communities are not equally vulnerable to these threats. Risk and harm are often disproportionate for marginalized communities or people in otherwise vulnerable positions. Therefore, safety often cannot be defined universally. Potential threats and harms will vary not only from platform to platform, but participant to participant.

As noted throughout the proposed signals, the focus is on elevating the positive, or outlining criteria for building public-friendly digital space: in this case, putting safeguards in place allows people to feel safe.

Related Concepts

There are a number of closely related concepts that help explain the landscape of online safety, including security and risk.

Though the terms safety and security are often used interchangeably, security can be seen as the process or actions taken to ensure safety. For example, locking your doors is a security measure taken to make you feel safe from someone entering your house. Similarly, in the context of online platforms, their security measures will determine whether participants on that platform are safe.

Additionally, though the concepts are related, safety should not be considered the *opposite* of risk. As noted by information science researcher Anthony Pinter and collaborators, because much of the literature around online safety (particularly as related to adolescents) equates harm to risk exposure, solutions are often focused on abstinence; if you disclose less information online or if you participate less online, you will be less exposed to threats and therefore be safer. However, harm is only a potential outcome of risk exposure. Therefore, reducing exposure to risk is not always the ideal solution for increasing safety, and can result both in decreasing the potential benefits of online interaction and in lessening the development of coping skills. Other strategies include threat mitigation and vulnerability reduction.

Why It's Important

Technologies, including online platforms, are typically not created with the intention of causing harm to the people who use them. However, harms experienced at both an individual and collective level are very real, causing people to feel unsafe online. Threats to safety, regardless of whether they are real or perceived, negatively impact the quality and enjoyment of people's online interactions. This feeling can often lead to withdrawal from online spaces. For example, a recent study of online abuse among women in Bangladesh, by Google researcher Nithya Sambasivan and collaborators, revealed that online safety is one of the largest barriers to gender-equitable technology use. The opposite of this, real and perceived safety, can help people have positive digital experiences.

Online harms such as privacy violations can lead to serious physical safety threats. When someone is "doxxed," their home address and other personal information is shared online, which can have severe harms in the real world. As an extreme example, "swatting" is when someone calls in a false tip to the police, resulting in a SWAT team being sent to someone's home; at least one individual has already been killed as a consequence. However, even when the safety threat is not physical, the psychological harm can be severe in cases of online harassment, where the content can be extremely offensive, derogatory, and violent. This type of content is harmful even for those who encounter it incidentally but are not the direct target.

Additionally, in the context of intimate partner violence, abusers often exploit technology for both surveillance and intimidation, including using social media platforms to stalk and harass their victims, as explored by security and privacy researcher Diana Freed and collaborators. Privacy violations can also lead to *reputational* as well as psychological harms. For example, both hacking and revenge porn have resulted in sexually explicit images being shared without consent; multiple such cases have resulted in the victims committing suicide, even when the images were fake.

Threats to online safety are also likely to further marginalize already vulnerable or marginalized groups, since participants in online communities are not equally vulnerable to these threats. Threats to safety, both offline and online, are amplified for, for example, women, people of color, and LGBTQ individuals. Some vulnerable communities are at increased risk for threats like harassment, or experience disproportionate harms from threats like hate speech. These vulnerabilities are also intersectional—that is, compounded by intersecting identities and structures of power—as detailed in legal scholar Kimberlé Crenshaw's theory of intersectionality.

How We Can Move the Needle

One of the reasons that safety can be so difficult to tackle for online platforms is that it often involves stakeholder tensions: for example, the tension between keeping children or adolescents safe and respecting their privacy and autonomy. Much of the solution-oriented research around child and adolescent online safety covers education to help children understand risks (such as games and storytelling, as explored by educational technology researcher Fotis Lazarinis and colleagues), or monitoring tools or parental control strategies (e.g., as explored by HCI researcher Pamela Wisniewski and collaborators). For the latter, some of the most promising work, such as that from HCI researcher Brenna McNally and collaborators, has involved youth as direct stakeholders, through participatory design methods focused on mitigating this tension. A number of researchers have concluded that strategies that emphasize digital literacy and communication that fosters trust may be more effective than pure restriction or monitoring.

As another example of a value tension, one commonly proposed solution for negative behavior online that harms others, like cyberbullying or hate speech, is to decrease anonymity online. However, though it is true that in some cases anonymity can increase rule-breaking behavior, in other cases anonymity is what keeps people safe. For example, research from social computing researcher Andrea Forte and colleagues showed that some Wikipedia users remain anonymous in order to mitigate risks of harassment, reputation loss, surveillance, or even violence. Similarly, many people who participate in support communities (e.g., LGBTQ youth) might not use their real names due to potential privacy and safety threats, as detailed by information science researcher Brianna Dym and colleagues. Therefore, not all platforms should have "real name" requirements.

Another tension for online safety intersects with both inclusion and free speech. Inclusion on an online platform means that a diversity of individuals feel comfortable expressing their views; however, it is also possible that some views actively make the space less safe for others. Debates about where to draw the line between free expression on the one hand and hate speech and harassment on the other are happening in the context of almost every social media platform. See our signal Invite Everyone to Participate for more on inclusion.

Despite these tensions, from a policy perspective, platforms have to make decisions about what kind of behavior and content to allow or disallow in order to keep participants safe. For example, many online platforms have explicit rules against many of the threats described here, like spam and harassment. However, content rules can be difficult to enforce, particularly at scale, which means that platforms typically rely on a combination of algorithmic and human content moderation, each of which has benefits and limitations. Social computing researcher Shagun Jhaver and collaborators have suggested that, for both paradigms, increased transparency in moderation decisions can both help newcomers learn rules and decrease the potential for future violations.

Another strategy for making sure that rules most effectively ensure people's safety is to give people a voice in rulemaking and/or consider social norms within a community when determining rules. Platforms might also allow sub-communities to create their own rules so that they can navigate their own safety concerns. For example, in addition to being bound by both content guidelines and "reddiquette" that include rules like "Remember the human," individual subreddits on Reddit can have their own rules that moderators may enforce as they see fit. As a result, many subreddits have specific rules tailored to their community around issues like harassment, inappropriate content, and hate speech. Social computing researchers

Casey Fiesler and Amy Bruckman have also shown that people are more likely to follow rules that reinforce social norms within a community than those that are imposed from the outside.

Of course, some communities or individuals are simply toxic, and can negatively impact an entire platform. Platforms might decide to ban certain content entirely (e.g., via hashtags, which as shown by social computing researcher Stevie Chancellor and colleagues, is not always effective) or even entire communities. When Reddit banned a number of subreddits for repeated violations of rules around hate speech and harassment, social computing researcher Eshwar Chandrasekharan and collaborators showed that this action successfully decreased this type of content on the platform overall. In contrast, as an example of a strategy for promoting safety without removing content, Tumblr displays an "Everything okay?" message of support with resources when users search for content related to self-harm.

Of course, rules can only go so far, and in the end, when platforms consider user personas, they must consider the bad actor. People with ill intent—predators, stalkers, harassers, racists—*will* be using that platform to harm other people. Platform designers need to anticipate these actions and potential harms, and design to mitigate them. This process should also include special consideration for vulnerabilities, which includes marginalized communities. Platform design has the potential to reinforce existing power structures by making it easier or more difficult for certain voices to be heard, as highlighted by Scheuerman and collaborators. For these reasons, safety concerns for the most vulnerable, even if they are not in the majority of the userbase, should be of high importance.

How to Measure

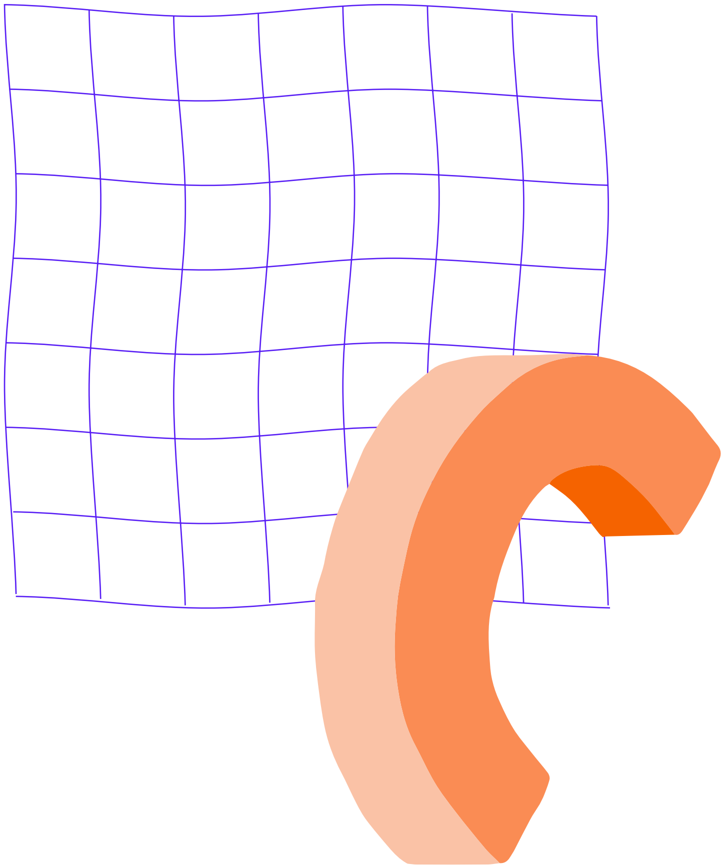
One traditional way of measuring safety in other domains, as explained by psychologist Rhona Flin and colleagues, has been "lagging indicators" such as number of accidents; however, there has been a growing movement towards more "leading indicators" such as safety audits. In the context of online platforms, audits might look similar to "threat modeling," which is a common practice in cybersecurity, and provides a method for determining potential threats and vulnerabilities. In considering how threat modeling can be used as a foundation for security requirements, computer scientist Suvda Myagmar and collaborators have suggested that to identify threats, analysts should ask questions like: who are my potential adversaries, what are their motivations, and what are their goals?

Similarly, one can ask what threats to any kind of safety might exist on this platform, particularly as instigated by potential bad actors. What vulnerabilities do users have? And what aspects of platform policy or design would eliminate these threats? Rather than waiting until after harm has occurred to measure it, such forward-thinking analysis might help prevent harm before it occurs.

Another potential indicator is how safe people *feel*, which can only be measured by asking people on the platform this question. One such method used by multiple researchers has been diary studies, in which users reflect at regular intervals on their experiences on the platform as related to risks and threats to safety. Leaving such reports open ended (as opposed to defining specific threats) will also allow a platform to build a taxonomy of perceived threats to safety, as security and privacy researcher Elissa M. Redmiles and colleagues have detailed.

Redmiles et al.'s diary study asked participants to detail safe and unsafe experiences on Facebook, and found that both tended to fall into general categories of privacy, security, and community. As another example, HCI researcher Pamela Wisniewski and collaborators' diary study asked teens to report weekly on risks they encountered online that fell into four types: information breaches, online harassment, sexual solicitation, and exposure to explicit content.

Though designers should be engaging in speculation about potential harms, getting frequent feedback from users about their real experiences of harms is likely the best way to measure the overall safety climate on a platform.



Foundational Works

- Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T., & Dell, N. (2018). **"A stalker's paradise": How intimate partner abusers exploit technology**. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1-13

- Hartikainen, H., Iivari, N., & Kinnula, M. (2016). **Should we design for control, trust or involvement? A discourses survey about children's online safety**. Proceedings of the 15th International Conference on Interaction Design and Children, 367-378.

- Redmiles, E. M., Bodford, J., & Blackwell, L. (2019). **"I just want to feel safe": A diary study of safety perceptions on social media**. Proceedings of the International AAAI Conference on Web and Social Media, 13, 405-416.

- Scheuerman, M. K., Branham, S. M., & Hamidi, F. (2018). **Safe spaces and safe places: Unpacking technology-mediated experiences of safety and harm with transgender people**. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 1-27.

- Wisniewski, P., Ghosh, A. K., Xu, H., Rosson, M. B., & Carroll, J. M. (2017). **Parental control vs. teen self-regulation: Is there a middle ground for mobile online safety?**. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 51-69.

Further Reading

- boyd, d., & Hargittai, E. (2013). **Connected and concerned: Variation in parents' online safety concerns**. Policy & Internet, 5(3), 245-269.

- Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016). **#thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities**. Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, 1201-1213.

- Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). **You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech**. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), 1-22.

- Crenshaw, K. (1990). **Mapping the margins: Intersectionality, identity politics, and violence against women of color**. Stanford Law Review, 43, 1241-1299.

- Dym, B., Brubaker, J. R., Fiesler, C., & Semaan, B. (2019). **"Coming out okay": Community narratives for LGBTQ identity recovery work**. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-28.

- Fiesler, C., & Bruckman, A. S. (2019). **Creativity, copyright, and close-knit communities: A case study of social norm formation and enforcement**. Proceedings of the ACM on Human-Computer Interaction, 3(GROUP), 1-24.

- Flin, R., Mearns, K., O'Connor, P., & Bryden, R. (2000). **Measuring safety climate: Identifying the common features**. Safety Science, 34(1-3), 177-192.

- Forte, A., Andalibi, N., & Greenstadt, R. (2017). **Privacy, anonymity, and perceived risk in open collaboration: A study of Tor users and Wikipedians**. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 1800-1811.

- Jhaver, S., Bruckman, A., & Gilbert, E. (2019). **Does transparency in moderation really matter? User behavior after content removal explanations on Reddit**. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-27.

- Lazarinis, F., Alexandri, K., Panagiotakopoulos, C., & Verykios, V. S. (2020). **Sensitizing young children on internet addiction and online safety risks through storytelling in a mobile application**. Education and Information Technologies, 25(1), 163-174.

- Magkos, E., Kleisiari, E., Chanias, P., & Giannakouris-Salalidis, V. (2014). **Parental control and children's internet safety: The good, the bad and the ugly**. 6th International Conference on Information Law and Ethics (ICIL 2014), 1-18.

- Myagmar, S., Lee, A. J., & Yurcik, W. (2005, August). **Threat modeling as a basis for security requirements**. In Symposium on Requirements Engineering for Information Security (SREIS) (Vol. 2005, pp. 1-8).

- McNally, B., Kumar, P., Hordatt, C., Mauriello, M. L., Naik, S., Norooz, L., ... & Druin, A. (2018). **Co-designing mobile online safety applications with children**. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1-9.

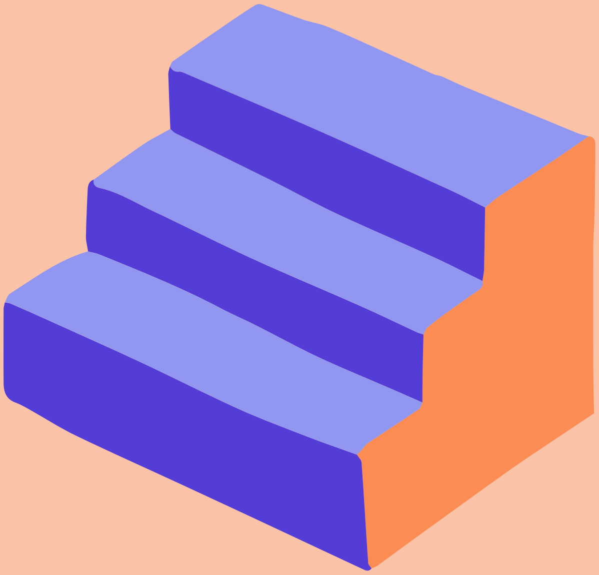
- Pater, J., & Mynatt, E. (2017). **Defining digital self-harm**. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 1501-1513).

- Pinter, A. T., Wisniewski, P. J., Xu, H., Rosson, M. B., & Caroll, J. M. (2017). **Adolescent online safety: Moving beyond formative evaluations to designing solutions for the future**. Proceedings of the 2017 Conference on Interaction Design and Children, 352-357.

- Sambasivan, N., Batool, A., Ahmed, N., Matthews, T., Thomas, K., Gaytán-Lugo, L. S., ... & Consolvo, S. (2019). **"They don't leave us alone anywhere we go": Gender and digital abuse in South Asia**. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-14.

- Wisniewski, P., Xu, H., Rosson, M. B., Perkins, D. F., & Carroll, J. M. (2016). **Dear diary: Teens reflect on their weekly online risk experiences**. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 3919-3930.

Expert Q&A



Three key questions with **Casey Fiesler**, University of Colorado Boulder

How does this principle help create a world we'd all want to live in?

The internet offers amazing potential for people to connect, learn, and support each other. However, threats to individual or community safety can overshadow or suppress the benefits. Frequent encounters with triggering content, harassment, privacy violations, disinformation, and hate speech (among others) will not only harm people at an individual level, but also communities at a collective level. Any measures that platforms can take to mitigate these harms and ensure people's safety on those platforms will result in better experiences, help people realize the benefits of online communities, and help to maintain an active userbase. Moreover, behavior online impacts the broader world, and minimizing problems like harassment and racism online also contributes to minimizing such harm more broadly.

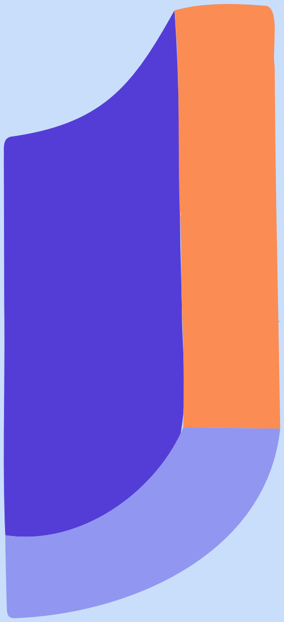
If you were to envisage the perfect social media, messaging or web search platform in terms of maximizing this principle, what would it look like?

Safety will look quite different on different platforms that have different userbases and purposes. However, in a general sense, safety should involve both discouraging bad behavior and encouraging pro-social behavior. This means a strong content moderation system where rules are contextual, clearly explained, transparent in terms of enforcement mechanisms, and determined with input from the community. Both policy and design can also encourage community members to look out for each other.

How would you measure a messaging, social media, or web search platform's progress against this principle?

Feelings of safety are highly subjective. Though there are some types of content that might be objectively detectable (e.g., hate speech), and platforms may find metrics that measure the presence of such content to be helpful, I think that engaging directly with users is likely to provide more contextually sensitive and nuanced information about threats and safety on a platform. Importantly, such user research should involve a diversity of users and communities, with a focus on those who are particularly vulnerable to harm. Conducting such research at regular intervals should provide metrics of progress, and also allow platforms to discern the effectiveness of changes made to policy or design.

Survey results



By Jay Jennings, Taeyoung Lee, Tamar Wilner, and Talia Stroud,
Center for Media Engagement

We conducted a survey with participants in 20 countries to understand more deeply how the signals resonated with people globally. Please find more about the methodology here.

The survey asked people to evaluate whether it was important for platforms to "ensure that people feel safe," and asked people to assess how well the platforms perform with respect to this signal. People were only asked about the platforms for which they are "superusers," by which we mean people who identify the platform as their most used social media, messaging, or search platform.

We analyzed how different demographic and political groups rate the importance of this signal, as well as the platforms' performance. In particular, we looked at age, gender, education, ideology, and country. We did this analysis for five platforms: Google, Facebook, YouTube, Facebook Messenger, and WhatsApp.[1] Only statistically significant results are shown and discussed.

[1] The analyses include only countries where at least 200 people responded that the social/ message/ search platform was the one that they use most frequently, and then only those platforms where we had data for at least 1,000 people. For Google, this includes all 20 countries. For Facebook, this includes 18 countries and excludes Japan and South Korea. For YouTube, this includes Brazil, Germany, Ireland, Japan, Malaysia, Singapore, South Africa, South Korea, and the United States. For Facebook Messenger, this includes Australia, Canada, France, Ireland, Norway, Poland, Romania, Sweden, the U.K., and the United States. For WhatsApp, this includes all countries except Canada, Japan, Norway, Poland, South Korea, Sweden, and the United States. Note that the total number of respondents varies by platform: Google = 19,554; Facebook = 10,268; YouTube = 2,937; Facebook Messenger = 4,729; and WhatsApp = 10,181. The larger the sample size, the smaller the effect that we are able to detect.

Importance of the Signal

We first examined whether platform superusers thought that the signal was important. For Facebook Messenger superusers in France, this was the most important of all 14 signals. It was second-most important for Messenger superusers in Norway and Romania; for WhatsApp superusers in Argentina, France, Mexico, and Romania; for Instagram superusers in Argentina; and for Facebook superusers in France and Romania.

Importance ranking: Ensure people's safety

A ranking of "1" means that the signal was seen as the most important of the 14 signals for superusers of a given platform in a given country based on a survey of over 20,000 people across 20 countries.

	Facebook	Youtube	Instagram	WhatsApp	FB Messenger	Google
Argentina	3		2	2		3
Australia	3	8		3	3	4
Brazil	10	11	11	7		7
Canada	5				3	5
France	2			2	1	3
Germany	11	11	10	4		6
Ireland	3	7		3	3	3
Italy	5			3		5
Japan		4				4
Malaysia	12	12	11	10		11
Mexico	3			2		3
Norway	3				2	5
Poland	11				8	10
Romania	2			2	2	5
Singapore	8	11		4		5
South Africa	10			5		5
South Korea		12				6
Sweden	7		6		5	5
UK	4			3	3	4
US	5	10			3	5

Signal most important

1
2
3
4
5
6
7
8
9
10
11
12
13
14

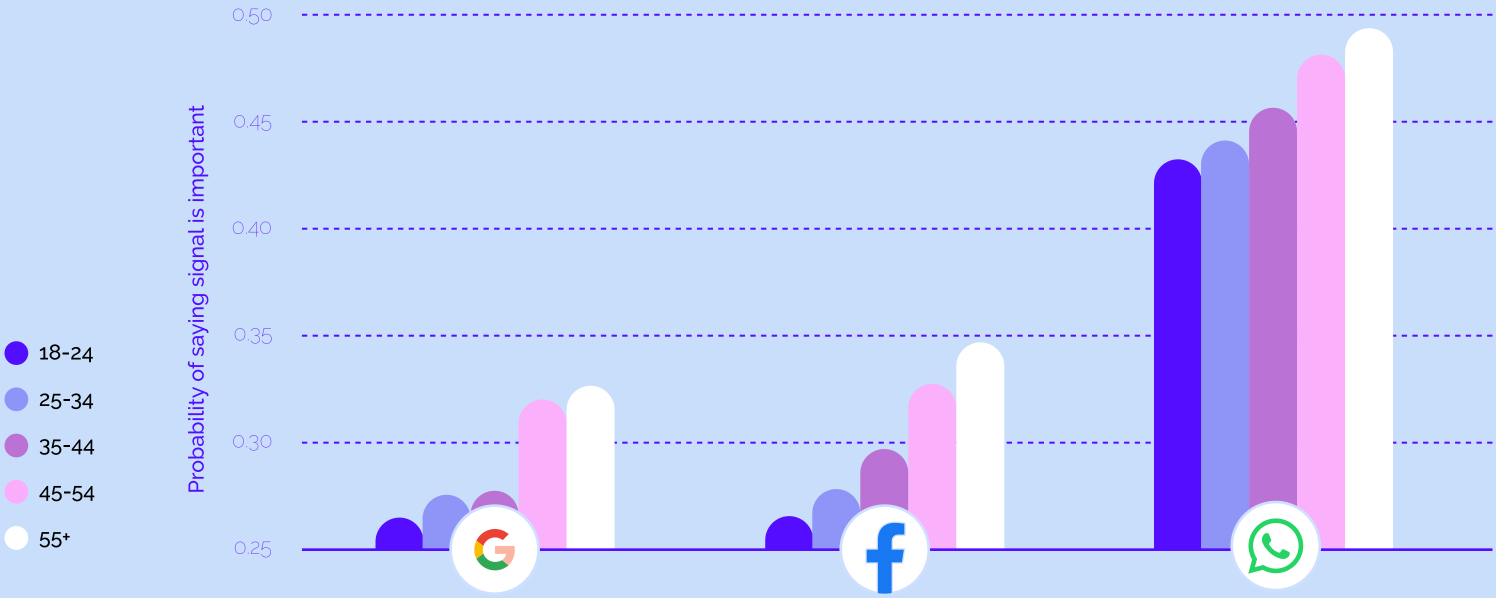
Signal least important

Data from the Center for Media Engagement. Weighted data. Asked of those who indicated that a given social media, messaging or search platform was their most used. Question wording: Which of the following do you think it is important for [INSERT SOCIAL, MESSAGING OR SEARCH PLATFORM] to do? Please select all that apply. Data only shown for those countries where at least 200 survey respondents said that the platform was their most used social media, messaging, or search platform.

Importance of the Signal by Age[2]

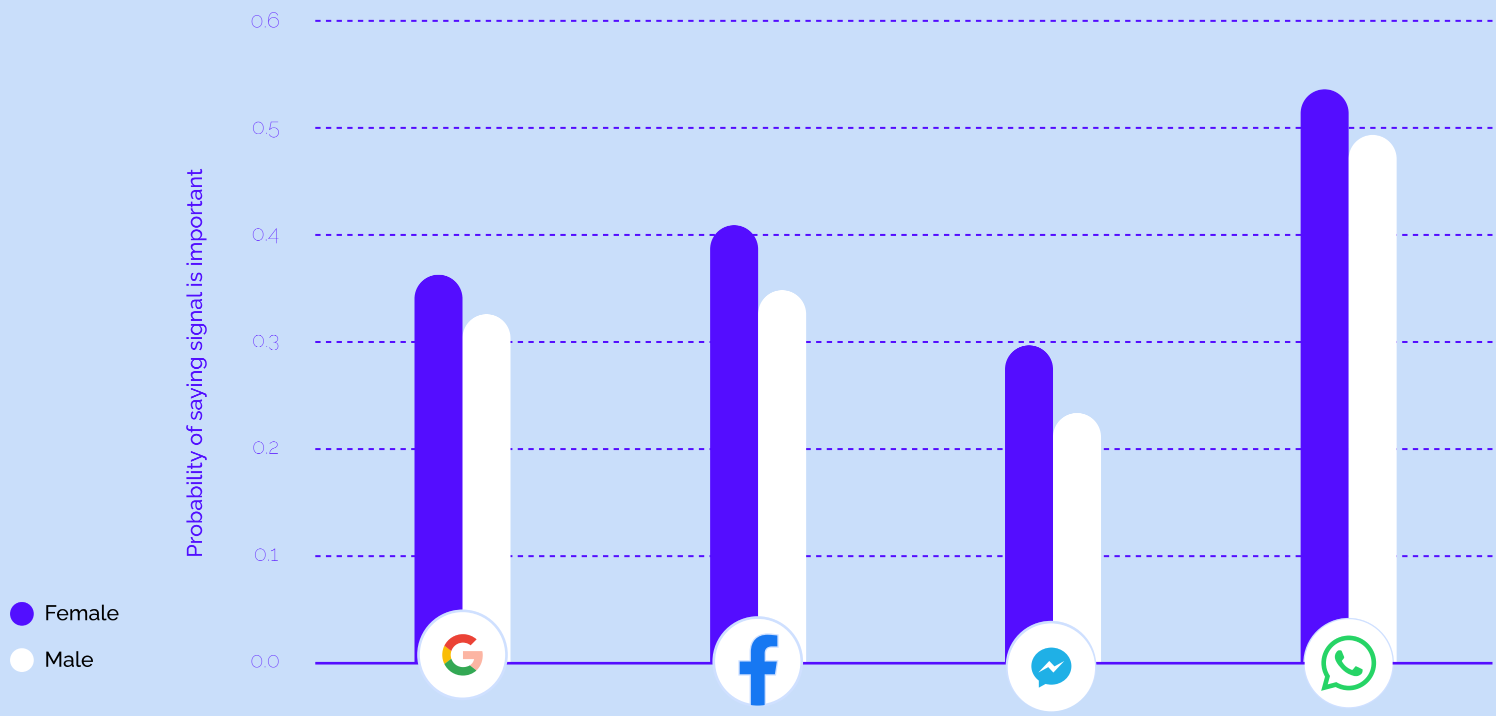
Age predicted whether people thought it was important to "ensure that people feel safe" for three platforms: Google, Facebook, and WhatsApp. For each of these platforms, those who were younger were less likely to say this signal was important and those in older age groups were more likely to say it was important.

[2] Results shown are predicted probabilities, calculated from a logistic regression analysis predicting that the signal is important based on age, gender, education, ideology, and country, each treated as a categorical variable. The baseline (based on the excluded categories) is a 55+ year old male with high education and middle ideology from the United States (except for WhatsApp, where the baseline is South Africa).



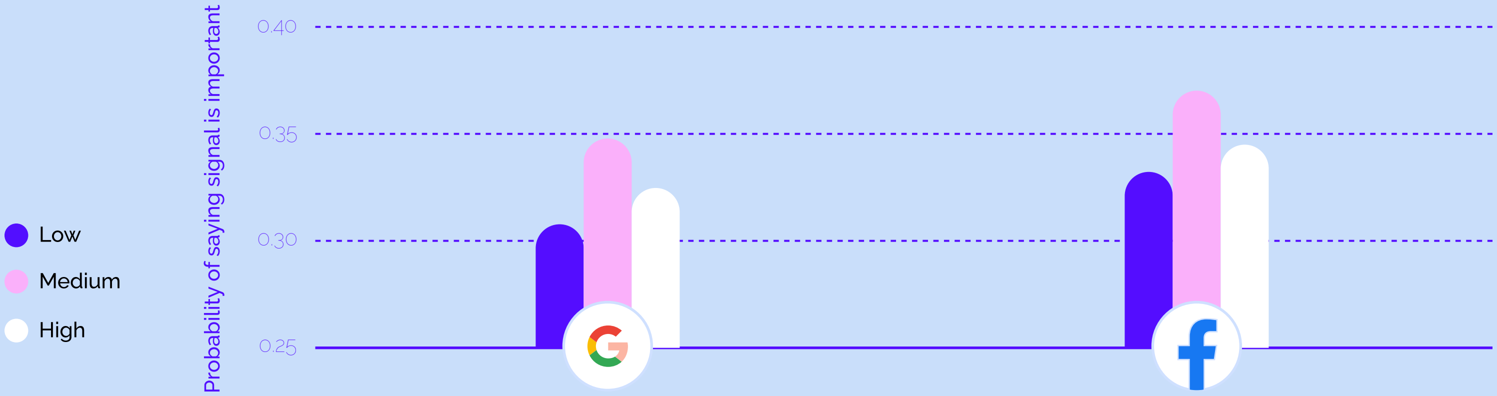
Importance of the Signal by Gender

For Google, Facebook, Facebook Messenger, and WhatsApp, men and women differed in the importance they ascribed to ensuring people's safety. Women were more likely than men to say this signal was important for all four of these platforms.



Importance of the Signal by Education

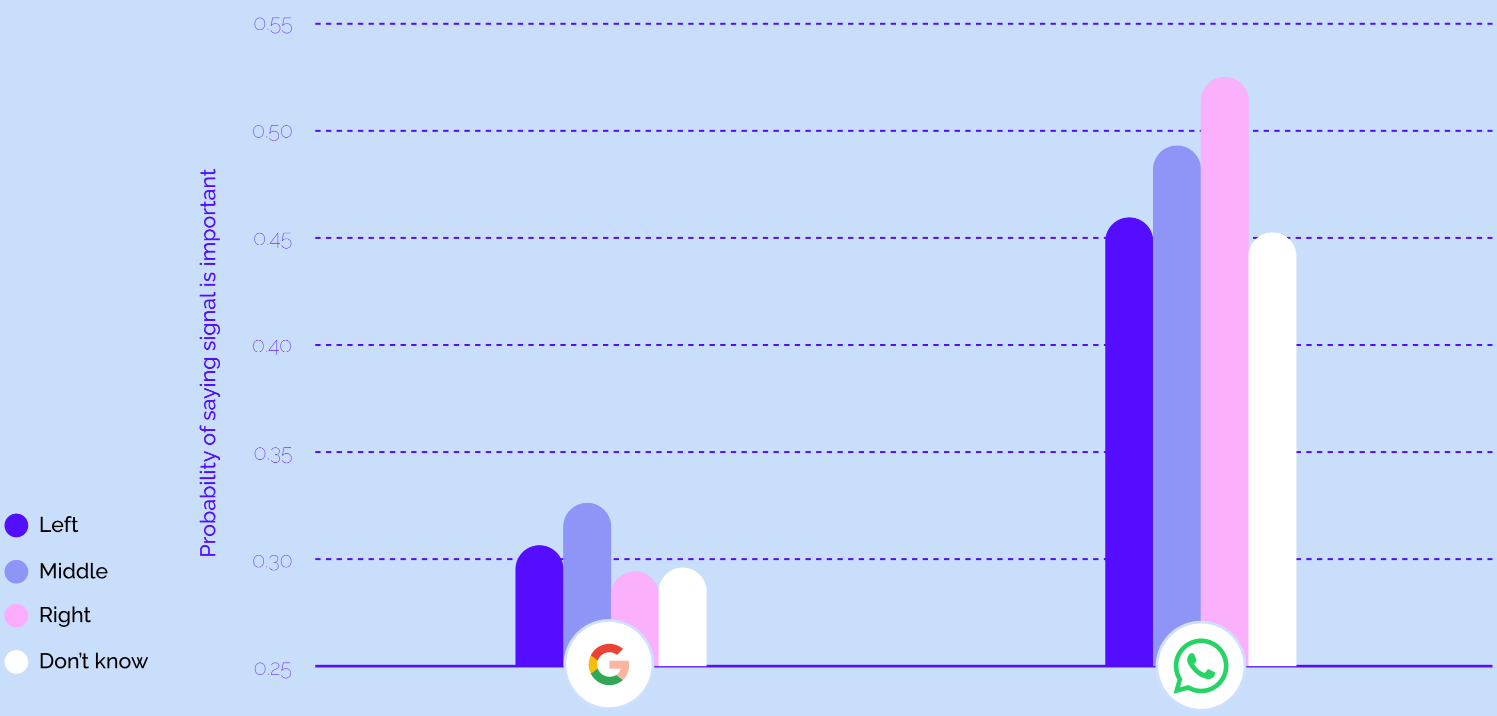
Respondents' view of the importance of ensuring people's safety differed by education levels for Google and Facebook. For both Google and Facebook, those with medium education levels were the most likely to say that ensuring people's safety was important.



Importance of the Signal by Ideology[3]

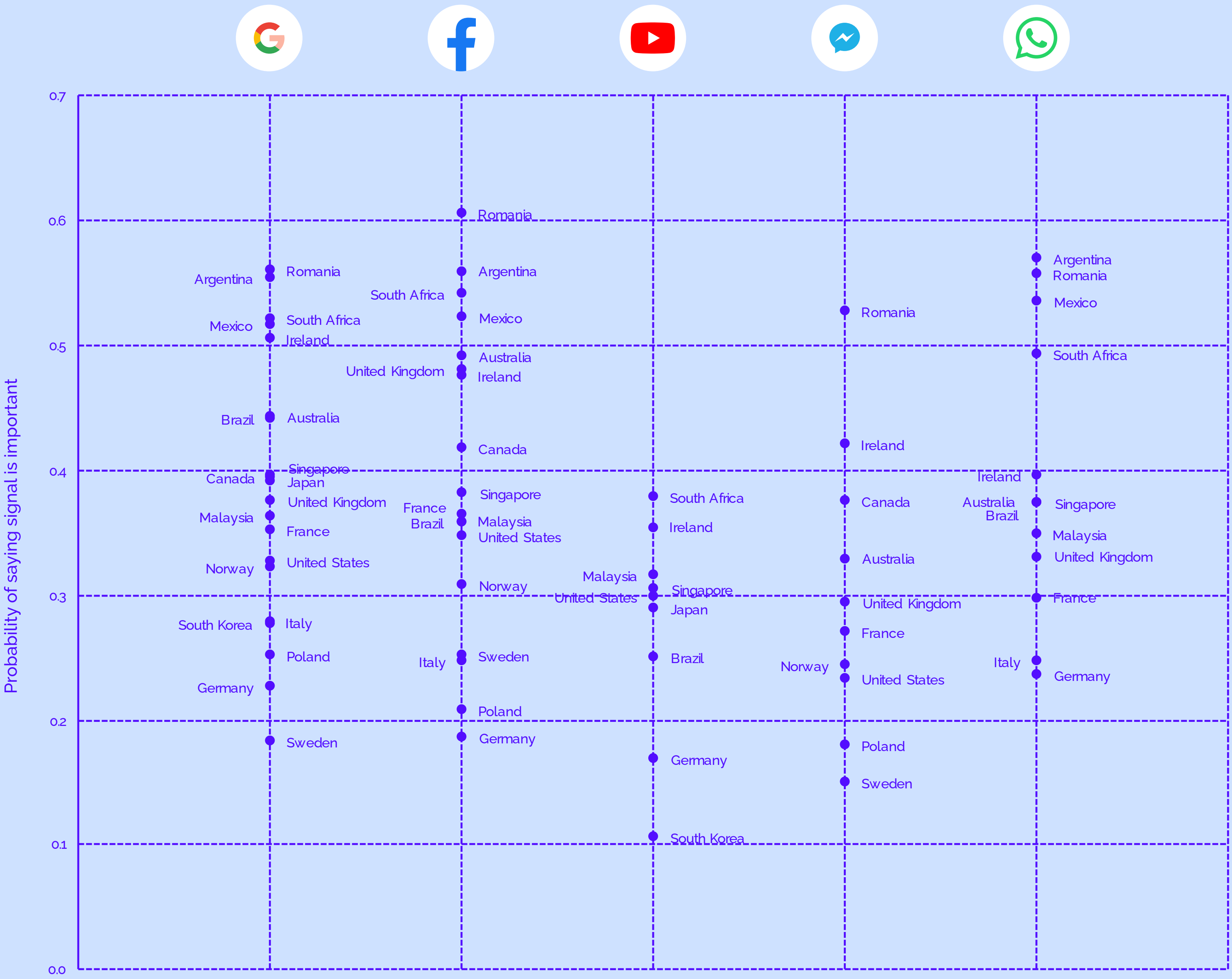
There were differences across political ideology in those who say it is important to "ensure that people feel safe" for Google and WhatsApp. For Google, those on the political right and those who didn't know their ideology were less likely to say this signal was important compared to those in the middle. For WhatsApp, those on the political right were more likely to say this signal was important compared to those with other ideologies and those in the middle were more likely to say that this signal was important compared to those who didn't know their ideology.

[3] Ideology was asked on a 10-point scale and people were given the option of saying "don't know." This was recoded into 4 categories (1 through 3, 4 through 7, 8 through 10, and "don't know").



Importance of the Signal by Country

There was significant variation by country for all five of the platforms we examined, based on how important people thought it was to “ensure that people feel safe.” The chart below shows the probability of saying that the signal is important by platform and by country. Overall, survey respondents in Romania and Argentina were the most likely to say this signal was important. Sweden and Germany were the least likely to say this was important.



Platform Performance on the Signal

For specific platforms, superusers were first asked to say on which of the signals they thought that the platform was doing well, and then on which of the signals they thought that the platform was doing poorly. We then categorized people's responses as (0) believe that the platform is doing poorly, (1) believe that the platform is doing neither well nor poorly, or (2) believe that the platform is doing well. Across countries, WhatsApp was rated as performing slightly better than neutral and Facebook slightly worse than neutral by superusers.

Performance index: Ensure people's safety

Responses of "2" indicate that everyone in a particular country thought that the platform was performing well on a signal; responses of "0" indicate that no one in a particular country thought that the platform was performing well on a signal based on a survey of over 20,000 people across 20 countries.

	Facebook	Youtube	Instagram	WhatsApp	FB Messenger	Google
Argentina	0.9		1.0	1.1		1.0
Australia	0.8	0.9		1.1	0.9	0.9
Brazil	1.1	1.0	1.1	1.1		1.1
Canada	0.8				1.0	1.0
France	0.9			1.1	0.9	1.0
Germany	0.9	0.9	0.9	0.9		0.9
Ireland	0.9	1.0		1.1	1.0	1.0
Italy	1.0			1.0		1.0
Japan		1.0				1.0
Malaysia	1.0	1.1	1.0	1.1		1.0
Mexico	0.9			1.1		1.0
Norway	0.9				1.0	0.9
Poland	0.9				1.0	1.0
Romania	0.9			1.2	1.0	1.0
Singapore	0.9	0.9		1.0		1.0
South Africa	0.9			1.2		1.0
South Korea		0.9				1.0
Sweden	0.8		0.9		0.9	0.9
UK	0.8			1.1	0.9	0.9
US	0.8	0.9			0.9	0.9

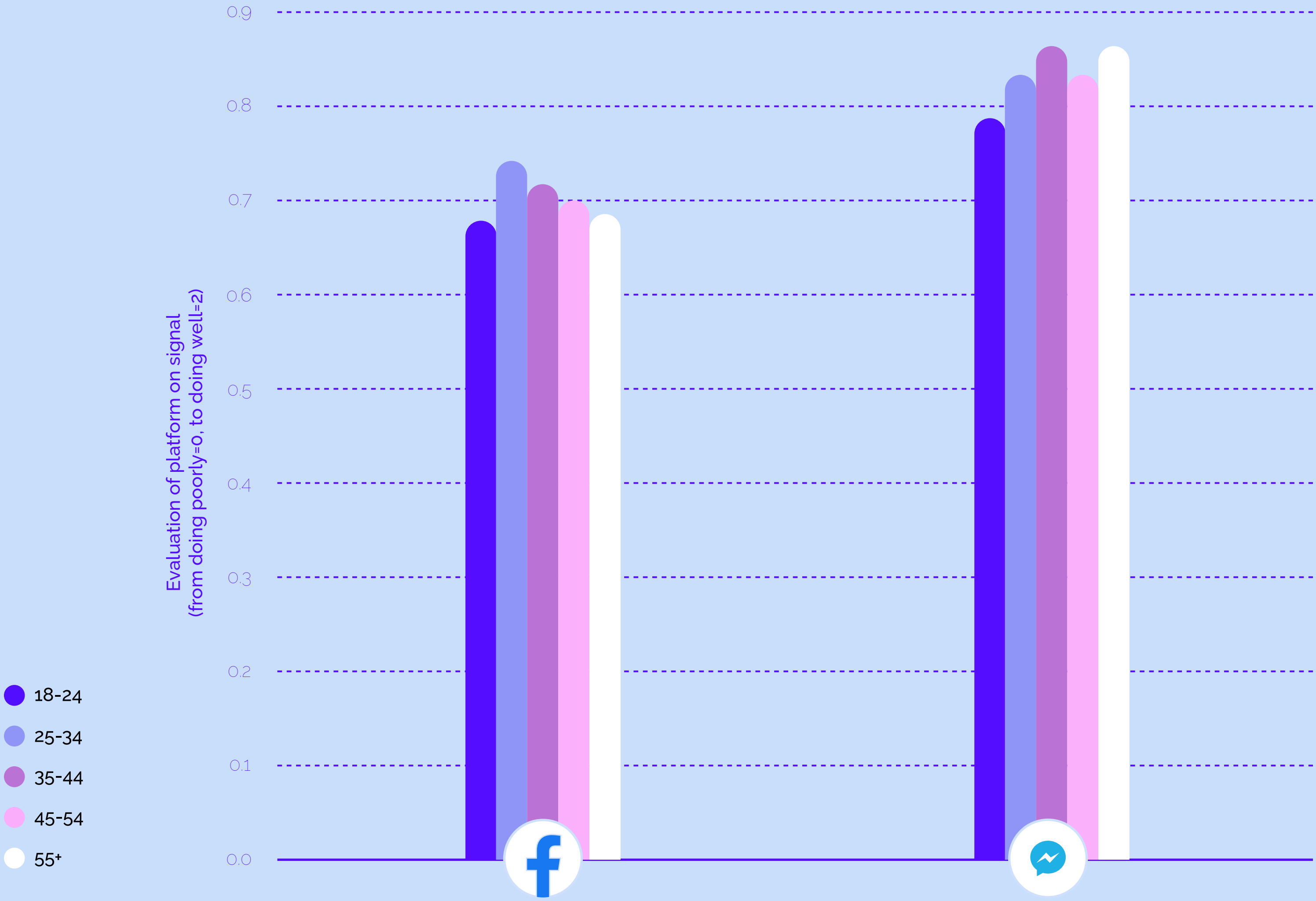
Well
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
Poor

Data from the Center for Media Engagement. Weighted data. Asked of those who indicated that a given social media, messaging or search platform was their most used. Question wording - Which of the following do you think [INSERT SOCIAL, MESSAGING OR SEARCH PLATFORM] does well at? Please select all that apply. And which of the following do you think [INSERT SOCIAL, MESSAGING OR SEARCH PLATFORM] does poorly at? Please select all that apply. Data only shown for those countries where at least 200 survey respondents said that the platform was their most used social media, messaging, or search platform.

Platform Performance on the Signal by Age[4]

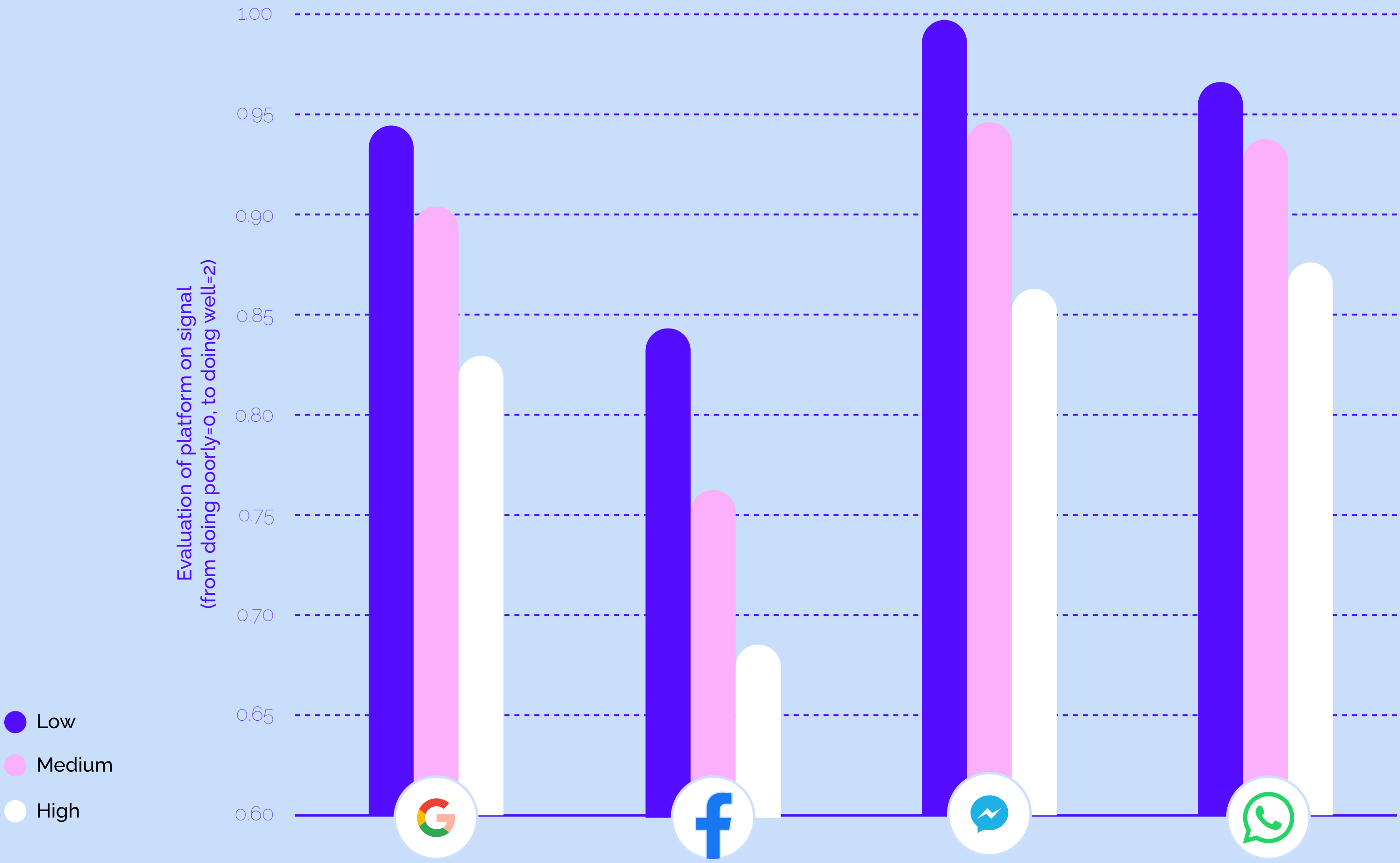
Only for Facebook and Facebook Messenger did the responses about performance in ensuring that people feel safe differ by age. Those in the youngest age group (18-24) gave the lowest ratings while those in the 25-34 age range gave the highest ratings of Facebook's performance in ensuring that people feel safe. For Facebook Messenger, the youngest age group gave lower ratings than those in the 35-44 and 55+ age groups.

[4] Results shown are predicted responses, calculated from a regression analysis predicting that the signal is important based on age, gender, education, ideology, and country, each treated as a categorical variable. The baseline (based on the excluded categories) is a 55+ year old male with high education and middle ideology from the United States (except for WhatsApp, where the baseline is Germany).



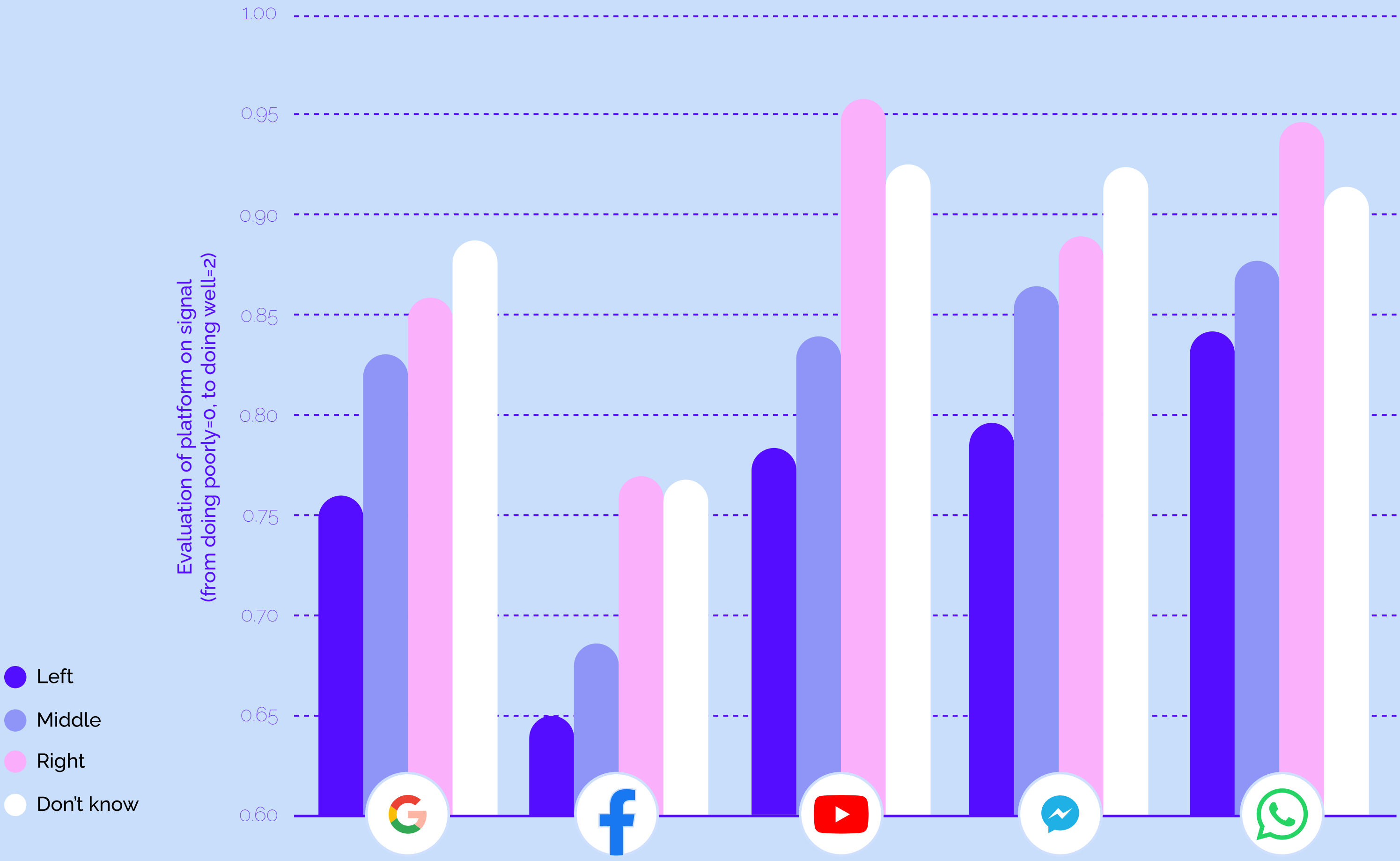
Platform Performance on the Signal by Education

For Google, Facebook, Facebook Messenger, and WhatsApp, ratings of the platforms' performance on "ensuring that people feel safe" differed by education levels. For all four of these platforms, those with lower education levels gave higher performance ratings and those with higher education levels gave lower performance ratings for ensuring people's safety.



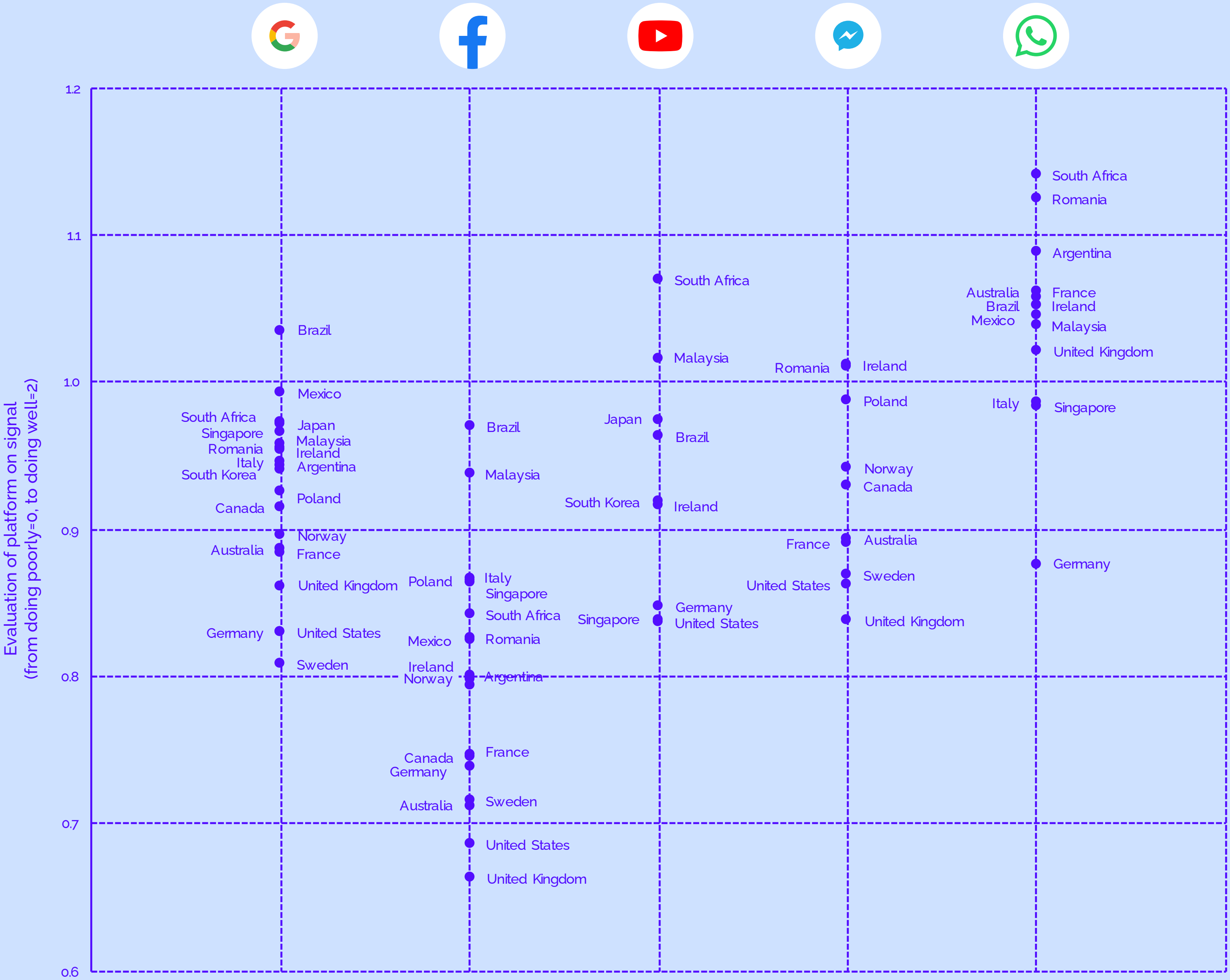
Platform Performance on the Signal by Ideology

For all five platforms, responses differed by political ideology for platform performance on ensuring that people feel safe. For Google and Facebook Messenger, those on the left evaluated the platform's performance more poorly than all of the other ideological groups and those who didn't know their ideology evaluated the platform better than those with middle ideologies. For YouTube and WhatsApp, those on the right and who didn't know their ideology believed that the platform was performing better with respect to this signal than those on the left or in the middle. For Facebook, those on the left evaluated the platform's performance more poorly than those on the right.

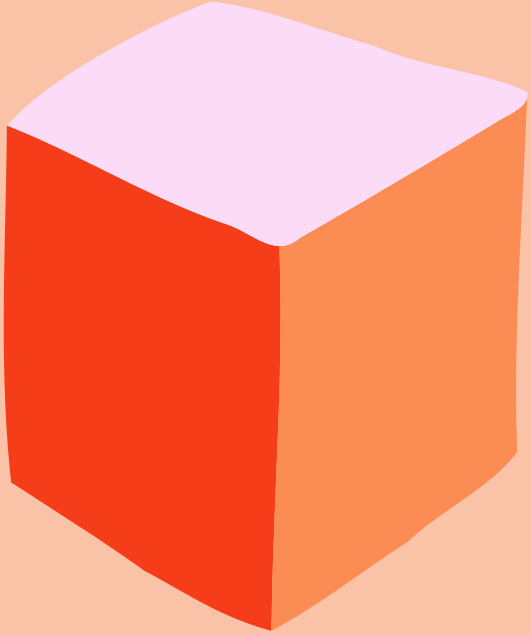


Platform Performance on the Signal by Country

There was variation by country in evaluations of platform performance. The chart below shows how superusers rated the platforms' performance in each country, controlling for age, gender, education, and ideology from "doing poorly" (0) to "doing well" (2). In general, those in South Africa and Brazil tended to say that the platforms performed well while those in the United States, United Kingdom, Germany, and Sweden thought they performed poorly.



Focus group report



By Gina Masullo, Ori Tenenboim, and Martin Riedl,
Center for Media Engagement

We conducted two focus groups in each of five countries (Brazil, Germany, Malaysia, South Africa, and the United States). Please find more about the methodology here. Participants were asked to reflect on their social media experiences and the proposed signals. With respect to this signal, participants made several observations. Please note that all names included are pseudonyms.

People were very concerned about the presence of pedophiles, stalkers, scammers, and others with bad intent on social media. They realized that they needed to protect themselves but also wanted platforms to do more.

> “
> *A few months back, my daughter was telling me, ‘I am going out.’ ‘With whom?’ ‘One of my friends from my online Twitter chat.’ ‘Do you know who that is? Do I know? Anyone from school?’ ‘No, this one is online.’ That is a big no-no. I was so shocked. It makes parents paranoid.” – Farah, Malaysian focus group participant*

“Social networks might bring people

> *“I find that legislation should be changed in a way that leads to a better security frame. That you must register with your real name and perhaps identify yourself with your ID card or so. So, I would wish for some more protection.” – Walter, German focus group participant*

together but can also pull them away, can bring evilness too, so we have to wise up… especially with kids,” said Miguel, of Brazil. In line with this sentiment, Farah, of Malaysia, talked about the challenge of protecting his daughter: “A few months back, my daughter was telling me, ‘I am going out.’ ‘With whom?’ ‘One of my friends from my online Twitter chat.’ ‘Do you know who that is? Do I know? Anyone from school?’ ‘No, this one is online.’ That is a big no-no. I was so shocked. It makes parents paranoid.”

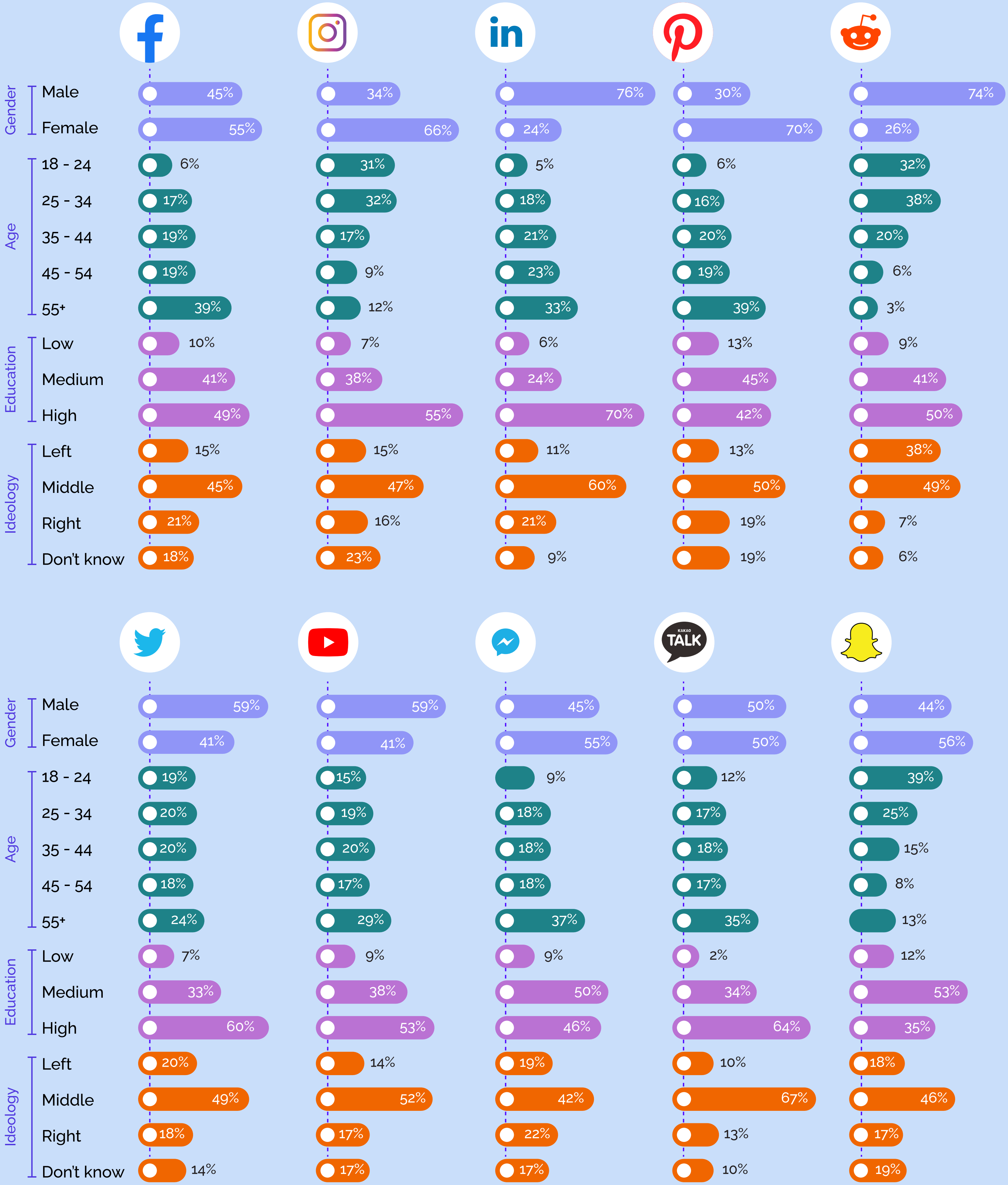
Tracy, of the U.S., called on users to police themselves on social media: “Don’t go posting other people’s addresses, harassing them or stalking them. Have a free voice but within certain reasonable parameters. Say what you want to say, but don’t go beyond that because then that’s not free speech.”

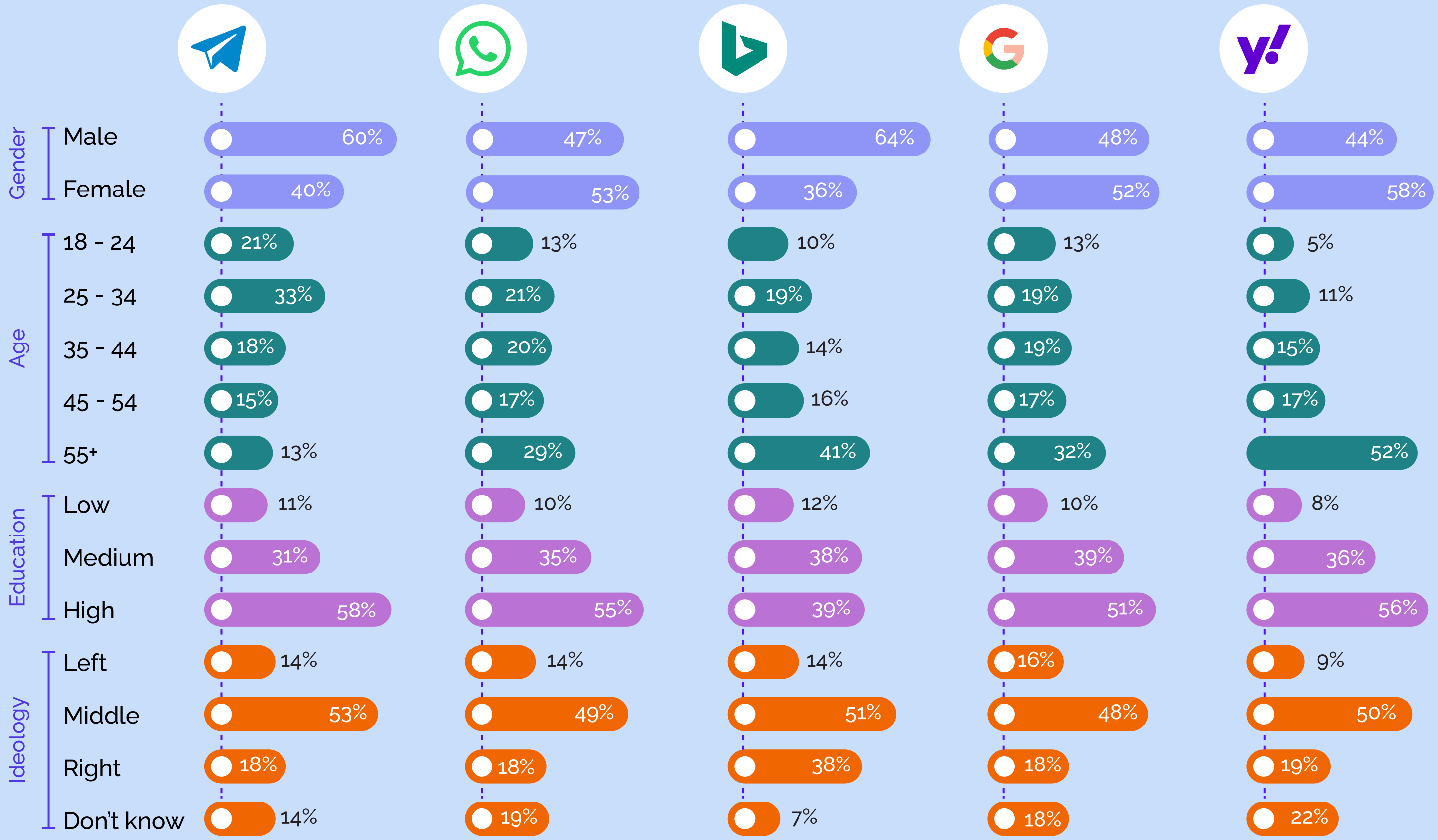
Other participants expected social media to do more to ensure users’ safety. Jéssica, of Brazil, felt platforms should do more to protect young people from pedophiles. “The owner of social media must establish the order, in order to be in a social medium, there must be limits. It’s not a madhouse,” she said.

Walter, of Germany, thought that social media should make it harder for people to have fake accounts. “What annoys me more and more,” he said, “is that I don’t need to register with my real name but just with a nickname. And then I can stir up hatred on the internet. This is something that increasingly annoys me. I find that legislation should be changed in a way that leads to a better security frame. That you must register with your real name and perhaps identify yourself with your ID card or so. So, I would wish for some more protection.”

User demographics from survey

Based on the survey respondents across all 20 countries, we looked at the demographics of superusers. For example, of those naming Facebook as their most used social media platform, 45% are male and 55% are female.




Gender
Male
Female
Age
18 - 24
25 - 34
35 - 44
45 - 54
55+
Education
Low
Medium
High
Ideology
Left
Middle
Right
Don't know
60%
40%
21%
33%
18%
15%
13%
11%
31%
58%
14%
53%
18%
14%
47%
53%
13%
21%
20%
17%
29%
10%
35%
55%
14%
49%
18%
19%
64%
36%
10%
19%
14%
16%
41%
12%
38%
39%
14%
51%
38%
7%
48%
52%
13%
19%
19%
17%
32%
10%
39%
51%
16%
48%
18%
18%
y!
44%
58%
5%
11%
15%
17%
52%
8%
36%
56%
9%
50%
19%
22%

Logo glossary

Social media

Facebook

Instagram

LinkedIn

Pinterest

Reddit

Twitter

YouTube

Messaging

Facebook Messenger

KakaoTalk

Snapchat

Telegram

WhatsApp

Search engines

Bing

Google

Yahoo

Civic Signals