

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm

RED FLAG NO.



6

RED FLAGS IN THE VALUE PROPOSITION

Such as:

- Social media, messaging and online platforms that allow individuals to post abusive content, form groups with the purpose of inciting hatred or violence, or engage in discriminatory practices.
- Platforms used by children and young people that allow users (including adults) to post videos and images of violent, sexual or dangerous behavior.
- Apps designed for use by specific groups (e.g., members of the LGBTQI community) that can where the risk of exposure of group membership gives rise to potential States surveillance and/or persecution.
- Online gaming sites offering chat rooms, where players may engage in misogynistic behavior, graphic language and imagery, and predatory child grooming and abuse.
- Online marketplaces that make it possible for individuals to refuse to do business – e.g., sell a service, exchange goods, offer jobs or rent property - with individuals of a certain ethnicity or sexual orientation.
- Adult websites that make it possible for individuals to upload videos or images of people without their consent, or illegal content such as of the sexual exploitation of women or children.
- Content moderation on platforms that restricts freedom of expression and disproportionately targets specific groups of people.
- Platforms that can be used for phishing scams.

Note: AI technologies can augment and potentially increase the risk of adverse human rights risks connected to the examples above (see Red Flags 5, 9 and 16).



HIGHER-RISK SECTORS:

- Social Media and Messaging Platforms
- Web-based calling and video services
- Online marketplaces and sharing economy platforms (such as online classified advertisements, dating, recruitment and real estate sites or applications)
- Platforms with child and/or youth users
- Online gaming sites and related chat rooms

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm

RED FLAG NO.



6

RED FLAGS IN THE VALUE PROPOSITION

HIGH-LEVEL
DECISION-MAKING

RISK TO PEOPLE

RISK TO THE
BUSINESS

UNGP's & SGD's
ANALYSIS

TAKING ACTION



KEY QUESTIONS FOR LEADERS TO ASK OR BE ASKED:

- How does the company assess whether its use of digital platforms is, or risks, enabling human rights harms? Does this include a review of how strategies to increase user numbers, user engagement and revenue may undermine the company's efforts to operate responsibly?
- How does the company seek to prevent the posting and spread of harmful content and/or misinformation? Does it engage in content moderation, and how does it strike a balance between preventing harmful content and potential censorship?
- What are the processes the company has in place to engage with civil society and other experts to remain aware of the potential adverse impacts on people from the use of digital platforms, and to explore any dilemmas that may arise in seeking to mitigate those risks?
- How does the company engage with peers and governments to mitigate potential human rights issues, e.g., defining industry standards and laws aimed at protecting against platform-related harms?
- If the platforms include AI-enabled functionality, how does the company ensure that AI solutions are developed and deployed with human oversight, transparency in algorithmic decision making, privacy protections and credible ways for people to raise grievances or seek remedy.

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm



6

RED FLAGS IN THE VALUE PROPOSITION



RISKS TO PEOPLE

Digital platforms can create meaningful connections and services — but they can also pose serious risks to people's rights, wellbeing, and safety. From harassment and exploitation to discrimination and criminal activity, these technologies can enable online and offline harm to people in the absence of appropriate governance, safeguards, and rights-respecting design principles.

AI further changes the nature of these risks by enabling harmful content to be generated, personalized and distributed at unprecedented speed and scale, while recommendation systems can rapidly amplify that content to users most likely to engage with it. The opacity of these systems can make it more difficult for users, regulators and companies to understand, challenge and remedy resulting harms.

Risks to Life, Safety and Security

Digital platforms can enable serious risks to people's safety and security, particularly where harmful content and interactions are created, shared or amplified at speed and at scale. (*Right to life; Right to liberty and security of person; Right to health*).

- Companies running online [classified ads platforms](#) have been accused of [facilitating human trafficking, including sexual exploitation of children](#).
- Some adult websites [allow for non-consensual](#) content to be posted. Some companies have been reluctant to (or refused to) take down content when requested by victims while [profiting from the harm through ad revenue](#). Online dating platforms have become a mainstay in modern society, but [recent meta-research suggests](#) that they can be catalysts for sexual violence and abuse both on and offline.

- Hate speech and other forms of harmful communication on platforms can incite violence, including attacks on targeted individuals or communities. UN Human rights investigators looking into the genocide in Myanmar confirmed that [social media had played a “determining role” in the spreading of hate speech and the campaign of ethnic cleansing](#) against the Rohingya minority. In addition to the “[determining role](#)” of social media in perpetuating violence in [Myanmar](#), civil society reports have also looked into the catalytic effect of social media in the conflict in northern [Ethiopia](#).
- Platforms can be used to [incite armed violence](#) offline or contribute to crimes against humanity, [recruit members of terrorist organizations](#), or target organizations, journalists or individuals working in humanitarian and peace building efforts.

Risks to Freedom of Expression and Spread of Disinformation leading to real-world harm

Platforms shape how information is created, shared and governed, creating risks both from the spread of harmful content and from the restriction of legitimate expression. Critics have called out [the tension](#) between removing extreme and harmful content, and the profitability of such ‘hyper-polarising’ content and its power for recruiting new users. [Attention is increasingly turning to platform design itself](#), with courts, regulators and the public scrutinizing whether features such as recommendation systems, ranking and engagement optimization foreseeably amplify harmful content and contribute to rights impacts. (*Right to freedom of opinion and expression; Right to freedom of thought, conscience and religion; Right to free and fair elections; core rights protected under International Humanitarian Law*).

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm

RED FLAG NO.



6

RED FLAGS IN THE VALUE PROPOSITION



RISKS TO PEOPLE

- Disinformation and 'fake news' spread through social media and communications platforms is increasingly seen as one of the most pressing societal challenges. Disinformation can cause harm to people and societies, for example [when it incites violence](#) toward particular groups. Offline harm can be triggered by the [use of deep fakes](#), by misinformation that leads to [voters lacking accurate information](#) about political candidates, and when people make [important decisions about their health](#) based on false information.
 - There is growing evidence of how some governments around the world are using [censorship and content manipulation to sway elections and restrict access to information](#). The [use of automated technology from private companies - including AI](#) - allows for large amounts of data on social media to be analyzed and aggregated in real-time and creates new possibilities for mass automated surveillance.
- There are also risks relating to **social media** in perpetuating discrimination and inciting violence against certain groups, or [fueling conflict](#) as compared to traditional media.
- At the same time (and illustrating the complexity of the issue), when content is wrongly moderated and removed from a platform, it can affect the users' [freedom of expression](#). TikTok was criticized for [filtering out videos from users](#) who did not meet certain aesthetic criteria.
- In recent years, social media companies such as [Meta and X](#) have removed third-party fact checking and shifted to community notes, which means that users collectively rate and fact check posts. Meta also recently revised its [Community Standards](#) on content moderation by lifting restrictions on certain topics, such as immigration and gender identity, which could lead to

increased risk of hate speech. However, [recent studies](#) have shown that a majority of social media users support content moderation and the removal of harmful social media content.

- Platforms like [Snapchat](#) pioneered the 'ephemeral post' feature (which was later adopted by Facebook and (then) Twitter), where messages and posts exist for only a certain period of time and then disappear 'forever'. While billed as a way to support more private modes of sharing, experts acknowledge the [added difficulty in monitoring and removing](#) toxic or harmful content from these more private interactions.

Risks to Privacy and Autonomy

The data-driven nature of platforms can expose users to risks relating to privacy, manipulation and loss of control over personal information. (*Right to privacy*).

- More and more people are also becoming victims of online and [social media scams and phishing](#). Scammers lure users into clicking [malicious links or ads](#), or [divulging personal information such as social insurance information or bank details](#) through fake ads, accounts or messages; or by connecting with users and deceiving them through the illusion of [romantic relationships](#).
- [Identity theft](#) is another concern through a range of modalities including criminals using information posted on social media, hacking into personal accounts taking them over, or through large scale data breaches from online platforms.
- There is growing evidence of how some governments are using social media platforms and associated data to conduct surveillance and restrict access to information, including through automated tools.

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm

RED FLAG NO.



6

RED FLAGS IN THE VALUE PROPOSITION



RISKS TO PEOPLE

- The [OHCHR](#) and [Freedom House](#) have long warned of the expanding use of digital platforms, behavioral data and AI-enabled surveillance tools by governments to monitor populations, manipulate information environments and restrict civic freedoms.
- The [emergence of AI bots operating independently](#) on various types of social media platforms, without any human oversight, is likely to introduce new avenues for harassment, hate speech and discrimination.

Risks to Equality and Non-Discrimination

Digital platforms were widely expected to reduce some forms of discrimination by standardizing transactions and reducing opportunities for interpersonal bias. While there is evidence that platforms have achieved these benefits in some contexts, experience has also shown that discrimination can persist or be exacerbated through platform design, algorithms, reputation systems and user interactions. In some cases, these features can institutionalize unequal treatment in ways that are less visible and more difficult to identify and remedy. (*Right to equality and non-discrimination*).

- In the rental housing market, [studies](#) have found that Black, Hispanic, and Asian hosts earn less, price lower, and receive fewer bookings than white hosts, resulting in higher drop-out rates. [Airbnb's internally released civil-rights audit](#) shows U.S. guests perceived as Black experienced lower booking acceptance rates compared to white guests. In Japan, real estate platforms that allow landlords or agents to indicate preferences

regarding foreign tenants - or that fail to remove discriminatory references in listings - can become connected to [discrimination](#) against non-Japanese residents.

- A [2016 study](#) found that drivers for ride sharing platforms Uber and Lyft were making black clients wait longer before accepting their trip requests and that drivers were more likely to cancel on people with “black-sounding” names.
- Job advertisements on job search platforms may contain discriminatory content specifying, for example, desired age or gender in the job post. Laws regarding discrimination in employment vary, such that postings that violate the right to non-discrimination may be legal in some jurisdictions.

Heightened Risks for Other Vulnerable Groups

Risks related to **children**:-

- The design of platforms used by young people (pre-teens, teenagers and young adults) can perpetuate the spread of videos and posts that reflect or promote harmful behavior, such as bullying, extreme dieting, [anorexia and body dysmorphia](#), [physical harm](#), drug use and inappropriate content such as [porn](#) and [suicide livestreams](#).
- Platforms designed to allow [users to chat without registration and to be randomly matched with strangers](#) pose particularly grave risks for children. Online gaming sites and their connected chat rooms for players have in some instances become [predatory grooming grounds](#) for child abuse. An issue that can be further exacerbated through the use of [virtual reality](#), where immersive experiences can expose users to new kinds of abuse.

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm

RED FLAG NO.



6

RED FLAGS IN THE VALUE PROPOSITION



RISKS TO PEOPLE

- Platforms that incorporate AI interaction systems - where users interact with bots - can be particularly harmful for young users, as articulated by a [2025 study by Common Sense Media](#), which asserted that “*Social AI companions pose unacceptable risks to teens and children under 18, including encouraging harmful behaviors, providing inappropriate content, and potential exacerbating mental health conditions.*” (See further Red Flag 8, Products that harm with higher levels of use).
- Subscription-based creator platforms, such as OnlyFans, have highlighted how business models built around direct creator-to-subscriber payments and limited content restrictions can increase risks of [child sexual exploitation and abuse](#). Recent reporting has raised concerns about the effectiveness of platform safeguards for preventing, detecting and responding to such harms.
- [AI-generated deepfake imagery](#) that facilitates the creation and sharing of non-consensual sexual content of women on poorly moderated platforms, leading to privacy violations, reputational harm, and emotional trauma.
- Location-based features can create heightened risks where privacy and security safeguards are inadequate. In 2024, [researchers found](#) that vulnerabilities in the design of several popular dating apps, including Grindr, could allow malicious users to pinpoint another user's location, exposing users to stalking, surveillance and other offline harms. These risks can be particularly acute for LGBTQI users in jurisdictions where they face discrimination, criminalization or violence, evidencing a need for [additional security protections](#) for members.

Labor Rights

Risks related to **Gender** and **Sexuality**:-

- Critiques have called out features on platforms which quantify competition and comparison metrics (likes, views and comments) as contributing to pressurized social interactions which can be harmful for health. [Addictive behaviors and mental health issues](#) can result from social media use for digital status seeking (looking for others to validate popularity online) and social comparison. [Negative mental health impacts](#) can be particularly acute among teenage girls.
- Platforms may fail to prevent or manage [gender-specific](#) or [racial](#) harassment and overtly [toxic interactions](#). Social media platforms can be used for practices like “doxing”, [swatting](#), [online hoaxes](#) and other forms of [harassment](#), including harassment that is discriminatory in nature.
- Platforms that use digital technologies to organize and mediate work through algorithmic matching are increasingly being [scrutinized](#) and criticized for [decent work deficits](#) – workers find themselves in competitive spaces where [high labour supply, increases in platform fees](#), etc. are resulting in low wages, under-employment and [precarious working conditions](#).
- [Working conditions for content moderators of social media](#) platforms such as Facebook, have also been reported as resulting in PTSD and mental health problems for workers. Already [marginalized members of communities in the global south report little access to mental health support](#) in the role. (See Red Flag 15, Business relationships with limited influence to address risk to people)

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm

RED FLAG NO.



6

RED FLAGS IN THE VALUE PROPOSITION



RISKS TO THE BUSINESS

Regulatory Risks

Despite their vast reach, social media platforms have been described as “[operat\[ing\] in a regulation-free zone](#)”, and are [increasing lobbying](#) efforts to maintain that status. Concerns about failures to meaningfully address impacts on people are leading, however, to calls for increased regulation (for example in [Australia](#), [France](#) and [Brazil](#)), with [debate](#) as to the form the regulation should take.

- Recent movements towards regulating platforms include the [UK Online Safety Act](#), which sets out strict guidelines governing the removal of illegal content and specific responsibilities with regard to children.
- The [EU Digital Services Act Package](#) (Digital Services and [Digital Markets Acts](#)) was announced by the European Commission in December 2020, aimed at ensuring a safe, rights-respecting online space in Europe, and a level-playing field for technology innovation and competitiveness across the region, and bolstered by substantial fines and penalties. The Acts entered into force in 2022, and the rules have been fully applicable since 2024.
- There has also been extensive [regulatory activity](#) – in [Australia](#), [France](#), [UK](#), [several US states](#), as well as other countries – to legislate social media use by and increase protection for children – such as social media age restrictions, and [parental consent requirements](#).
- Governments have begun imposing [direct operational constraints](#) on platforms where risks to children are judged to be insufficiently managed; for example, under Australia’s Online Safety Amendment (Social Media Minimum Age) Act 2024, platforms including Instagram, TikTok and YouTube must take reasonable steps to prevent users under 16 from creating or maintaining accounts, with potential penalties

of up to approximately AUD \$49.5 million for systemic non-compliance, requiring companies to deactivate accounts and implement age-verification systems.

- In 2026, the [UK government has announced plans](#) to address the ways in which social media and other online activities can exacerbate or incite violence against women. As an initial step they’ve notified technology companies that any non-consensual intimate image that is flagged must be taken down within 48 hours and automatically deleted if it is reuploaded. They have also committed to designating creating or sharing such images as a priority offence under the UK’s Online Safety Act.
- In July 2026, the European Commission [preliminarily found that TikTok breached the EU’s Digital Services Act](#) by failing to ensure sufficiently protective default account settings for minors, which could lead to a requirement for design changes and/or expose the company to a fine of up to 6% of its global annual turnover.

Reputation and Legal Risks

Online platforms have seen legal challenges, boycotts and widely disseminated online campaigns.

- March 2026 New Mexico verdict against Meta: A New Mexico jury [ordered Meta to pay \\$375 million](#) after finding the company misled users about the safety of its platforms and knowingly exposed children to sexual exploitation and mental health harms. Internal documents presented at trial suggested that Meta’s engagement-maximizing business model - designed to attract and retain young users - generated incentives that amplified the very risks the company publicly denied.

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm

RED FLAG NO.



6

RED FLAGS IN THE VALUE PROPOSITION



RISKS TO THE BUSINESS

- Meta has faced sustained legal, regulatory and governance scrutiny over allegations that Facebook's content recommendation and moderation systems contributed to serious human rights harms. CEO Mark Zuckerberg has been questioned in high-profile US Senate hearings about the platform's possible [role in inciting violence against the Rohingya in Myanmar](#); in an interim submission to the U.N. Human Rights Council, the fact-finding mission chair emphasized the "[determining role](#)" of social media networks in the conflict. Rohingya refugees also brought a US\$150 billion lawsuit alleging that Facebook amplified hate speech that contributed to the atrocities against the Rohingya. Although the [Ninth Circuit upheld dismissal of the case in 2026 on the basis of Section 230 of the Communications Decency Act](#), the litigation illustrates the significant legal, reputational and governance risks companies may face where platform design is alleged to contribute to severe human rights harms. Meta has also continued to face governance scrutiny, including a [2025 SEC whistleblower complaint](#) raising concerns about changes to the company's content moderation policies.
- In the aftermath of a December 2020 [New York Times expose](#), Pornhub took swift steps to change the platform's policies and user features, in reported contrast to its [lack of action](#) relating to calls for similar changes from a wide range of survivors and civil society organizations.
- In 2024, twelve school boards in Canada and a group of private [schools launched a lawsuit against Snapchat, TikTok, Instagram and Facebook worth \\$8 billion for impacts](#) on children's mental health, learning outcomes and wellbeing.
- In 2025, the [U.S. Federal Trade Commission issued orders](#) to seven companies (such as Meta, Alphabet, Open AI) that provide consumer-facing AI-powered chatbots seeking information on how these firms measure, test, and monitor potentially negative impacts of this technology on children and teens.
- Airbnb [faced several lawsuits](#) and a [social media campaign](#) regarding the racial discrimination experienced by users (see 'Risks to People' above). In 2017, Airbnb and the state of California reached an agreement that will [allow the state to conduct racial discrimination audits](#).
- Online gaming platform Roblox has faced several lawsuits, including [a lawsuit brought by a state attorney general](#) that alleges failures to protect children from predators and exposure to harmful content. Changes in policy and processes have resulted (see 'Mitigation' below).

Business Opportunity Risks

Similar trends are also being seen with social media, as platforms such as [Mastodon](#) and [Bluesky experienced exponential growth](#) as users sought them out for greater protection of privacy, security and ethical data practices. [Advertisers boycotted X](#) (formerly twitter) resulting in falling advertising revenues due to changes in content moderation policies after the takeover of the platform by billionaire Elon Musk; however, in 2024 [Musk sued the advertising agencies and companies for violating US anti-trust laws](#).

- Online chat platform Omegle was permanently shutdown in November 2023; following a '[product liability lawsuit](#)' by a victim; this is the first time a technology platform is being put on trial for the way it was built arguing that there is a

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm

RED FLAG NO.



6

RED FLAGS IN THE VALUE PROPOSITION



RISKS TO THE BUSINESS

fundamental defect in the design, which leads to harm.

- Brands may seek to avoid association with platforms perceived as contributing to harms to people. In 2021, [cosmetic brand Lush withdrew from Instagram, TikTok and Facebook](#), citing concerns about algorithm-driven harms and unsafe online environments, while focusing more heavily on owned channels and community engagement. While Lush retained activity on some other platforms, including YouTube, the move demonstrated how concerns about impacts on people can shape brand engagement decisions. A broader shift by advertisers and brands away from certain platforms could ultimately affect platform revenues.
- The range of potential human rights impacts, highlighted here and under Business Model Red Flag 16, that can occur under a single business model have been used to justify (including from [governments](#)) [anti-trust measures to break-up monopolistic platforms](#).

Financial - stock price - Risk

Social media companies, such as [Meta](#) and [Reddit](#), have experienced stock price impacts on several occasions in the past following content moderation related controversies. While downward pressure on stock prices in recent examples has tended to be shorter-term the potential longer-term impacts on reputation (and corresponding brand value) have not been fully evaluated.

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm

RED FLAG NO.



6

RED FLAGS IN THE VALUE PROPOSITION



WHAT THE UN GUIDING PRINCIPLES SAY:

**For an explanation of how companies can be involved in human rights impacts, and their related responsibilities, see [here](#).*

A company operating an online platform can cause human rights harms when it takes or fails to take a decision that results in people being prevented from enjoying rights such as the right to privacy, right to information, freedom of expression or their right to be forgotten. Examples include where a platform filters out user content or closes user accounts erroneously, or when a major data breach occurs that violates user privacy.

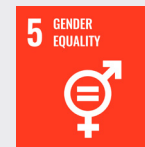
Companies operating online platforms can also contribute to a range of human rights harms when the design and functionality of platforms facilitates, third parties engaging in harmful behavior, including by failing to act on credible evidence of harm. In this context, harms might be experienced by:

- a user due to their own use or misuse of the platform,
- a user because another actor has used, misused or abused the platform,
- a third party due to how a user has used, misused or abused the platform.



POSSIBLE CONTRIBUTIONS TO THE SDGs:

Addressing impacts to people associated with this red flag indicator can positively contribute to a range of SDGs depending on the impact concerned, for example:



SDG 5: Achieve gender equality and empower all women and girls

- **Target 5.1:** “End all forms of discrimination against all women and girls everywhere.”



SDG 16: Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels

- **Target 16.1:** Significantly reduce all forms of violence and related death rates everywhere.
- **Target 16.2:** End abuse, exploitation, trafficking and all forms of violence against and torture including of children.
- **Target 16.10:** Ensure public access to information and protect fundamental freedoms, in accordance with national legislation and international agreements.

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm

RED FLAG NO.



6

RED FLAGS IN THE VALUE PROPOSITION

DUE DILIGENCE LINES OF INQUIRY:

- How does the company identify, assess and address risks that its platform, products or services may be used to cause or facilitate harm to people?
- How has the company engaged with potentially affected stakeholders, including vulnerable groups, to understand and address human rights risks associated with the platform?
- Does the company provide clear user standards or terms governing prohibited conduct and misuse of its platform?
- What systems are in place to detect, assess and respond to harmful, exploitative, illegal or otherwise abusive activity occurring through the platform?
- How does the company enable users or third parties in all markets to report harmful activity, and how does it respond to such reports?
- Does the company provide or participate in effective grievance mechanisms for individuals and communities adversely affected by its platform?
- How does the company measure the effectiveness of its efforts to prevent and address human rights impacts? What metrics, testing and governance processes are used?
- If deploying AI systems for content moderation, recommendations, advertising or other user-facing functions, how does the company ensure appropriate transparency, human oversight, testing and privacy protections?
- How does the company ensure transparency around key decisions affecting users, including content moderation, account restrictions, data access or automated decision-making?
- How does the company evaluate and respond to government requests for content removal or user information, including safeguards against unlawful censorship or restrictions on freedom of expression?
- Does the company know who is providing content moderation or trust and safety services, and how does it ensure appropriate working conditions, including mental health protections where workers may be exposed to harmful material?

Spotlight on Social Media and User-Generated Content Platforms

The following questions are particularly relevant where the platform hosts, recommends or amplifies user-generated content.

- Does the company make clear that discriminatory, abusive or exploitative behaviour will not be tolerated, including through user agreements and community standards?
- What content moderation policies, governance processes and technical systems are in place to prevent the rapid spread of harmful or abusive content?
- What measures are in place to prevent advertising associated with crimes such as child sexual exploitation or trafficking, including collaboration with relevant authorities?
- What measures are in place to ensure age-appropriate experiences and protections for children and young users?
- How does the company educate users about acceptable behaviour and platform rules?
- What systems are in place to identify and respond to discriminatory behaviour, harassment, non-consensual intimate imagery or other abusive interactions between users?
- What support is provided to employees and contractors engaged in content moderation, including counselling and mental health services?

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm

RED FLAG NO.



6

RED FLAGS IN THE VALUE PROPOSITION



MITIGATION EXAMPLES:

** Mitigation examples are current or historical examples for reference, but do not offer insight into their relative maturity or effectiveness.*

Online Platforms:

- In the run up to the 2020 US elections, Facebook (now Meta Platforms Inc.) announced a [range of steps](#) they were taking to ensure the integrity of the elections including by removing misinformation, violence-inciting posts, the creation of a Voting Information Center, the development of a new hate speech policy, as well as political advertising blackout periods the week before and after the election. However, in January 2025, the company announced that it was [removing third-party fact checking](#), and switching to an approach of 'community notes', meaning relying on users to flag and fact-check posts.
- Social media companies have been developing stronger moderation systems to flag, escalate and make decisions about discriminatory or otherwise abusive behavior (e.g., employing monitoring staff that are trained on the local context; convening groups of experts to monitor important topics, especially where hate speech or fake news can lead to serious harm). For example, in the past, Facebook (now Meta) introduced [use of AI](#) to limit the spread of hate speech and [improve the speed](#) of its removal and, with others including Twitter (now X), joined the [global pledge to fight hate speech online](#). As stated above, it needs to be noted that these companies have since then rolled back their fact-checking efforts.
- Content moderation: monitoring and removing content is, in principle, a viable risk mitigation strategy and many social media companies employ moderators to manage the related risks to people. However, a number of additional risks to people are inherent to this work (1) privacy risks related to having your content, personal information and private interactions monitored (2) censorship if companies make inappropriate or incorrect decisions and (3) risk to the mental health of the content moderators who are regularly exposed to harmful, toxic and violent content.
- Meta created an Independent Oversight Board to [take final and binding decisions on whether specific content should be allowed or removed from Facebook and Instagram](#). The Board considers content referred to it by both users and Meta and makes [recommendations](#). Members contract directly with the Oversight Board and are not Meta employees.
- Online recruitment companies, such as LinkedIn, use a ["multitude of tools and systems to proactively monitor content and identify activity that may be in violation of our policies"](#), deploying human reviewers where users identify and report discriminatory content in job postings.
- Online dating apps, such as Grindr, have [taken steps to implement additional security measures](#) that provide

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm

RED FLAG NO.



6

RED FLAGS IN THE VALUE PROPOSITION



MITIGATION EXAMPLES:

- enhanced protection to users, particularly important in countries where same-sex sexual activity is illegal.
 - Since 2018, twelve major companies—including Apple, Facebook (Meta), Google, Reddit, Twitter, and Github—have endorsed the [Santa Clara Principles](#) and the overall number of companies providing transparency and procedural safeguards has increased.
 - In 2019, in response to criticism of social media impacts on health, Instagram (owned by Meta) began a two-year [pilot testing feeds without like metrics](#) as a way of “depressuring Instagram” and to enable users to focus more on connection as opposed to competition. In 2021, the application rolled this out as an optional feature, allowing all users to be able to disable “likes” metrics.
 - Tiktok rolled [out new features to protect kids and teens in 2025](#), including greater parental controls and oversights, as well as a new “wind down feature” for teens on the app after 10pm – the aim being for ‘positive nudges’ to help teens build more balanced digital habits. However, the effectiveness of measures such as [parental controls have had mixed results](#), with parents reportedly struggling with their complexity, not using them, or children finding ways to circumvent them.
 - In response to criticism from regulators and researchers and following litigation concerning risks to children and adolescents, Meta’s [Instagram platform introduced “Teen Accounts”](#) with default safety restrictions and content-filtering features, first rolled out in the United States, United Kingdom, Canada and Australia, with subsequent expansion to the European Union and other market. These measures include automatically placing under-18 users into restricted accounts, limiting who can contact them, and blocking certain categories of age-inappropriate content.
 - The Roblox Corporation announced in November 2025 that the platform will roll out facial [age checks for its approximately 151 million users and an optional, age-based chat feature](#), so minor users will only be allowed to chat with other minors who are similar in age, and not adult users.
- Online Marketplaces:**
- Following high profile claims of discrimination from hosts, Airbnb “[hired a task force made up of high-profile civil rights activists](#)” and, in 2018, [announced](#) that hosts will not be able to see the pictures of guests until after the host has decided whether or not to accept the guest. The company introduced a [non-discrimination policy](#) outlining when a host may or may not reject an applicant and setting out Airbnb’s rights to suspend a host for repeated violations. It has also outlined an approach to [encouraging more hosts from communities of color in the United States](#).

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm

RED FLAG NO.



6

RED FLAGS IN THE VALUE PROPOSITION



MITIGATION EXAMPLES:

- Rideshare companies Uber and Lyft have [removed information that indicated a rider's gender and race from initial ride requests](#), “removing bias from the ride request stage” of the transaction. (However 2020 [research suggests](#) that bias against nonwhite and LGBTQ+ riders persists in the form of ride cancellations after confirmation).



ALTERNATIVE MODELS:

- As a response to potential risks, [rival service providers targeting groups discriminated against on a popular online housing rental platform](#) have emerged, with the value proposition of offering services with greater inclusivity, or [enhanced privacy](#) and data ownership.

HIGH-LEVEL
DECISION-MAKING

RISK TO PEOPLE

RISK TO THE
BUSINESS

UNGP's & SGD's
ANALYSIS

TAKING ACTION

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm

RED FLAG NO.



6

RED FLAGS IN THE VALUE PROPOSITION



OTHER TOOLS AND RESOURCES:

- The Danish Institute for Human Rights, (2021) [Human rights impact assessment of digital activities](#)
 - Global Network Initiative resources:
 - [Implementation Guidelines](#).
 - [Public Assessment Reports on Independent Assessments of GNI Members](#).
 - [Country Legal Frameworks Resource](#) – Providing information on privacy protections in national legal frameworks
 - [Across the Stack Tool](#) – Aimed at providing an overview of roles and responsibilities of actors across the technology stack, in order to support due diligence efforts.
- [UN Global Digital Compact](#).
- UNDP (2025) Human Development Report – [A matter of choice: People and possibilities in the age of AI](#)
- OHCHR B-Tech Project
- IHRB (2025), [Unpacking the human rights consequences of social media moderation backsliding](#)
- Article 19 (2023), [Content moderation and freedom of expression handbook](#)
- Council of Europe (2025), [Regulating content moderation on social media to safeguard freedom of expression](#)
- European Union Agency for Fundamental Rights (2023), [Online content moderation - Current challenges in detecting hate speech](#)
- OHCHR (2021) [Moderating online content: fighting harm or silencing dissent?](#)
- BSR (2021), A Human Rights-Based Approach to Content Governance, https://www.bsr.org/reports/A_Human_Rights-Based_Approach_to_Content_Governance.pdf
- UNICEF, [Keeping children safe online](#)
- European Parliament (2025), [Protecting children online Selected EU, national and regional laws and initiatives](#), [https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/769570/EPRS_BRI\(2025\)769570_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/769570/EPRS_BRI(2025)769570_EN.pdf)
- Fast Company (2020), What data experiments tell us about racial discrimination on Airbnb”, : <https://www.fastcompany.com/90460723/airbnbwhileblack-the-inside-story-of-airbnbs-racism-problem>
- European Commission against Racism and Intolerance, Combating Hate Speech, General Policy Recommendation : <https://rm.coe.int/ecri-general-policy-recommendation-no-15-key-topics-combating-hate-spe/16808b763a>

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Digital platforms that connect people or transmit content, where design choices can enable or amplify online and offline harm

RED FLAG NO.



6

RED FLAGS IN THE VALUE PROPOSITION



OTHER TOOLS AND RESOURCES:

- OpenCanada (2019), Bringing human rights standards to content moderation on social media: <https://www.opencanada.org/features/bringing-human-rights-standards-content-moderation-social-media/>
- Human Rights Watch (2018), Social Media's Moral Reckoning: <https://www.hrw.org/news/2018/12/21/social-medias-moral-reckoning>
- Investor Alliance for Human Rights, Information, Communications and Technology (ICT) Briefing: Discrimination: https://investorsforhumanrights.org/sites/default/files/attachments/2019-12/Investor%20Alliance%20Briefing_%20Discrimination.pdf
- New America (2020), Getting to the Source of Infodemics: It's the Business Model <https://www.newamerica.org/oti/reports/getting-to-the-source-of-infodemics-its-the-business-model/>
- New America (2020), It's Not Just the Content, It's the Business Model: Democracy's Online Speech Challenge: <https://www.newamerica.org/oti/reports/its-not-just-content-its-business-model/>
- The Conversation (2019), Digital Colonialism: [*Why Some Countries Want To Take Control Of Their Peoples Data From Big Tech*](#)

This resource is part of Shift's collection of Business Model Red Flags, developed as part of the Valuing Respect Project and generously funded by Ministry of Foreign Affairs Finland, the Norwegian Ministry of Foreign Affairs, and Norges Bank Investment Management. Learn more at: shiftproject.org/valuing-respect