THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Providing online platforms for individuals to interact where use of the platform can lead to harm to human rights

RED FLAG NO.

6

RED FLAGS IN THE VALUE PROPOSITION

Such as:

- Social media, messaging and online platforms through which individuals may post abusive content, form groups with the purpose of inciting hatred or violence, or engage in discriminatory practices
- Platforms predominantly used by children and young people that allow users (including adults) to post videos and images of violent, sexual or dangerous behavior
- Applications designed for use by specific groups that can increase the possibility of States surveilling and persecuting individuals from those groups (e.g. members of the LGBTQI community)
- Online gaming sites where players may use related chat rooms to engage in misogynistic behavior, graphic language and imagery, and predatory child grooming and abuse
- Online marketplaces through which individuals can 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 to which individuals can upload videos or images of people without their consent, or illegal content such as of the sexual exploitation of children

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)
- Platforms with high numbers of users being children and young people
- Online gaming sites and related chat rooms
- Cloud and hosting services companies offering the infrastructural backbone and computing power to businesses listed above

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Providing online platforms for individuals to interact where use of the platform can lead to harm to human rights

RED FLAG NO.

6

RED FLAGS IN THE VALUE PROPOSITION

KEY QUESTIONS FOR LEADERS TO ASK OR BE ASKED:

- How does the company assess whether its platform 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 prevent the posting and spread of harmful content? Does it enable users or third parties in all markets to report harmful or abusive content and how does it respond to such reports?
- Does the company have processes in place to engage with civil society and other experts to remain aware of the potential impacts on people of their platforms, and to explore any dilemmas that may arise in seeking to mitigate those risks?
- Is the company engaging with peers and governments to help define industry standards and laws aimed at protecting against platform-related harms?

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Providing online platforms for individuals to interact where use of the platform can lead to harm to human rights

RED FLAG NO.

6

RED FLAGS IN THE VALUE PROPOSITION

RISKS TO PEOPLE

Hate Speech, Harassment and Illegal Content
(*Right to equality and non-discrimination; Right to life, liberty and security; Right to freedom of thought, conscience and religion; Right to Just and favorable conditions of work; Right to highest attainable standard of physical and mental health)*:

- Online classified ads platforms have been accused in recent years of facilitating human trafficking, including sexual exploitation of children.
- Hate speech and other forms of harmful communication 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.
- Social media platforms can be used for practices like "doxing" and harassment, particularly where platforms fail to prevent or manage gender-specific harassment and overtly toxic interactions.
- 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.
- Companies can facilitate the harmful practices of online platforms through provision of "back-end" services, such as building custom tools to support platforms that impact rights.

Mis-/Disinformation and Censorship
(Right to freedom of opinion and expression; Right to freedom of thought, conscience and religion; Right to free and fair elections):

- 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, such as when it incites violence or discrimination against certain groups, when it targets the reputation of individuals, when voters lack accurate information about political candidates, or when people make important decisions about their health based on false information.
- Critics of these platforms call out the tension between removing extreme and harmful content, and the profitability of such "hyper-polarising" content and its power for recruiting new users.
- 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. Tik Tok was criticized for filtering out videos from users who did not meet certain aesthetic criteria.

"Ephemeral Post" Features that may exacerbate harm
(Right to Privacy; Right to freedom of opinion and expression; Right to equality and non-discrimination)

- Platforms like Snapchat pioneered the "ephemeral post" feature (followed by Facebook and 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

RISKS TO PEOPLE

sharing, experts acknowledge the added difficulty in monitoring and removing toxic or harmful content from more private interactions such as these.

Adverse Impacts on High-Risk Vulnerable Groups
(Right to Privacy; Right to highest attainable standard of physical and mental health; Right to Education):

- Platforms predominantly used by young people (pre-teens, teenagers and young adults) may allow videos and posts that reflect or promote harmful behavior, such as bullying, extreme dieting, anorexia, drug use, body dysmorphia, and inappropriate content such as porn and suicide livestreams.

- Platforms can expose young people to high-levels of targeted advertising and marketing with critics highlighting the inherent tension between advertising-based models that moderate content based on viewer engagement and content safety issues.

- Online gaming sites and their connected chat rooms for players have in some instances become predatory grooming grounds for child abuse.

- Dating platforms for the LGBTQI communities are vulnerable to data hacking and surveillance, and require additional security protections for their members.

Right to Equality and Non-discrimination
The introduction of technological platforms for transactions was expected by many to reduce or remove the inherent bias that can negatively affect the way that humans approach and conduct transactions with others. However, high profile studies and incidents have shown that discriminatory conduct has made its way into platform-based transactions, and in some cases, been exacerbated by platforms that institutionalize the discrimination.

- In the rental housing market, landlords offering rooms for accommodation who refuse to host on the grounds of assumed ethnicity or gender identity have been identified in various studies. In Japan, real estate platforms that a) allow landlords to select "Foreigner accepted/not accepted" or b) do not remove such references by landlords, can become connected to discrimination against non-Japanese. In the US, a A Harvard Business School study noted that, *"applications [to Airbnb] from guests with distinctively African-American names are 16% less likely to be accepted relative to identical guests with distinctively White names."*

- Similarly, photos and names were implicated in a 2016 study, that found that drivers for ride sharing platforms Uber and Lyft were found to make 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.

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Providing online platforms for individuals to interact where use of the platform can lead to harm to human rights

RED FLAG NO.

6

RED FLAGS IN THE VALUE PROPOSITION

RISKS TO THE BUSINESS

- **Regulatory and Legal Risks:** Despite their vast reach, social media platforms have been described as "operat[ing] in a regulation-free zone," and increasing lobbying efforts to maintain that status. Concerns about impacts on people are leading, however, to calls for increased regulation, including from some platforms themselves, with debate as to the form the regulation should take.
  - Recent movements towards regulating platforms include the upcoming UK Online Harms Bill, which will set out strict guidelines governing the removal of illegal content and setting out 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.

- **Reputation and Legal Risks:** Online platforms linked to discriminatory practices or content have seen legal challenges, boycotts and widely disseminated online campaigns.
  - For example, Airbnb faced several lawsuits and a social media campaign regarding the racial discrimination experienced by users set out above (see above "Risks to People"). In 2017, Airbnb and the state of California reached an agreement that will allow the state to conduct racial discrimination audits.
  - Facebook CEO, Mark Zuckerberg, faced high profile scrutiny in US Senate hearings, among other fora, when he was questioned about the company's possible role in inciting the genocide in Myanmar. In 2020, after expressing concern about Facebook's oversight efforts ahead of the US election, an ad-hoc group of high profile activists and academics convened a "Real Facebook Oversight Board", a "self-appointed proxy for the official Facebook Oversight Board."
  - 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 stark contrast to its lack of action relating to calls for similar changes from a wide range of survivors and civil society organizations.

- **Business Opportunity Risks:** Rival service providers targeting groups discriminated against on a popular online platform have emerged, with the value proposition of offering services with greater inclusivity, or enhanced privacy and data ownership.

- **Stock Price Risk:** Company stock prices, such as Facebook, Alphabet and Salesforce.com, are vulnerable to losses in the wake of human rights-related scandals and controversies.

- **Operational Risks:** The range of potential human rights impacts, highlighted in this document and under the Data Red Flag, that can occur under a single business model have been used to justify (including from governments) for anti-trust measures to break-up monopolistic platforms.

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Providing online platforms for individuals to interact where use of the platform can lead to harm to human rights

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 or incentivizes third parties to engage in harmful behaviour. 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:

Providing online platforms for individuals to interact where use of the platform can lead to harm to human rights

RED FLAG NO.

6

RED FLAGS IN THE VALUE PROPOSITION

DUE DILIGENCE LINES OF INQUIRY:

- How do we identify, assess and address discriminatory or otherwise abusive behaviors on platforms? Have we engaged with potentially vulnerable groups to educate ourselves on how our processes can be improved to combat discrimination or otherwise abusive content by other users?
- Do we make clear to platform users that discrimination or otherwise abusive behavior will not be tolerated? Have we incorporated this into user agreements? Do we have in place clear and detailed content moderation policies and processes to prevent viral spreading of discriminatory or otherwise abusive content?
- Do we have counseling programs in place for employed content moderators, regularly exposed to harmful, explicit or distressing online content?
- What are we doing to educate our users on what kind of content will and will not be tolerated on our platform?
- What systems are in place to ensure discriminatory behavior or exploitative, non-consensual or otherwise abusive content or interaction are flagged and managed (e.g. removed or otherwise dealt with)?
- What systems are in place to ensure that ads tied to crimes such as sexual exploitation, including of children, are prevented and dealt with, including through collaboration with the relevant authorities?
- What measures do we take to ensure only age-appropriate content is served to our young users?
- How do we track the effectiveness of our efforts to combat discrimination or other human rights impacts associated with our platform? What are the tests and metrics used?
- Do we provide or participate in effective grievance mechanisms that are accessible to individuals and communities at risk of discrimination by our platforms?
- Do we ensure transparency of processes, specifically with making user data available or with regard to content removal?

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Providing online platforms for individuals to interact where use of the platform can lead to harm to human rights

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 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.

- 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, Facebook has announced the use of AI to limit the spread of hate speech and improve the speed of its removal and, with others including Twitter, has joined the global pledge to fight hate speech online.
  - 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.

- Facebook and Twitter have created lead roles for human rights experts, and Facebook has reportedly commenced "making sure that people with human rights training are in the meetings where executives sign off on new product features." Facebook has also 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 Facebook. Members contract directly with the Oversight Board, are not Facebook employees and cannot be removed by Facebook.

- 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 [their] policies," deploying human reviewers where users identify and report discriminatory content in job postings.

THE BUSINESS'S COMMERCIAL SUCCESS SUBSTANTIALLY DEPENDS UPON:

Providing online platforms for individuals to interact where use of the platform can lead to harm to human rights

RED FLAG NO.

6

RED FLAGS IN THE VALUE PROPOSITION

MITIGATION EXAMPLES:

- Online dating apps, such as Grindr, have taken steps to implement additional security measures that provide enhanced protection to users, particularly important in countries where same-sex sexual activity is illegal.

Online Marketplaces:

- Follwing 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.
- 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 non-white and LGBTQ+ riders persists in the form of ride cancellations after confirmation).

Providing online platforms for individuals to interact where use of the platform can lead to harm to human rights

OTHER TOOLS AND RESOURCES:

- Fast Company (2020), *What data experiments tell us about racial discrimination on Airbnb*.
- European Commission against Racism and Intolerance, *Combating Hate Speech, General Policy Recommendation*.
- OpenCanada (2019), *Bringing human rights standards to content moderation on social media*.
- Human Rights Watch (2018), *Social Media's Moral Reckoning*.
- Investor Alliance for Human Rights, Information, Communications and Technology (ICT) Briefing: Discrimination.
- New America (2020), *Getting to the Source of Infodemics: It's the Business Model*.
- New America (2020), *It's Not Just the Content, It's the Business Model: Democracy's Online Speech Challenge*.
- The Atlantic (2016), *Facebook And The New Colonialism*.
- 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