RANKING DIGITAL RIGHTS

CONSULTATION DRAFT
Human rights risk scenarios: Algorithms, machine learning and automated decision-making

Note: This document is the second in a series of three documents that are being shared for consultation. The purpose is to obtain expert and stakeholder feedback on the concepts, principles, and standards for company best practice that will in turn inform the development of new indicators for possible inclusion in future iterations of the Ranking Digital Rights (RDR) Corporate Accountability Index. This document should be read after the *Rationale for RDR's methodology expansion to address algorithms, machine learning, and automated decision-making*, as it builds on concepts summarized in that overview. This document in turn should be read before the list of *Best Practices for Business and Human Rights: Algorithms, Machine Learning, and Automated Decision-Making*, which are based on the *Risk Scenarios* outlined below.

All documents can be downloaded from the RDR website at: https://rankingdigitalrights.org/methodology-development/2021-revisions/#algorithms.

What are human rights risk scenarios?

This document presents a number of different human rights risk scenarios, which are short narratives linking company practices to violations of human rights enumerated in the Universal Declaration of Human Rights (UDHR). These scenarios are directly derived from news reports or published research, and illustrate the human rights harms related to privacy and expression that can result from internet, mobile, and telecommunications companies' use of algorithms, machine learning, and automated decision-making. Mapping these scenarios enables us to identify recommended company practices that would either prevent or mitigate the risk and severity of these harms (as articulated in the *Best Practices* document), which are in turn used as the basis for developing research indicators to evaluate company disclosures of relevant policies and practices.

Each scenario relates to either the right to privacy (UDHR art. 12), the right to freedom of expression and information (UDHR art. 19), or both. Many also touch on other human rights, as further explained in the next section titled "Scope of Human Rights Harms." In some scenarios the company is the immediate cause of the violation, while in other cases the use of algorithms, machine learning, and automated decision-making, or companies' sale of such tools to third parties, facilitates human rights violations committed by other parties. In yet other scenarios, these technologies have a fundamental impact on the information environments of entire societies, in turn creating the enabling conditions for human rights harms. The scenarios are therefore organized into four categories, some of which overlap and intersect with one another, as detailed in the section "Types of human rights harms" below.

Taken as a whole, these risk scenarios highlight how the use of algorithms, machine learning and automated decision-making by internet, mobile, and telecommunications companies can threaten human rights, and provide the basis for "best practices" that would mitigate the risk and severity of these harms or provide remedy when such harms occur. *Best Practices* are presented in a companion document, to be read next.

Scope of human rights harms

RDR's current Index methodology focuses on identifying and mitigating individual human rights harms—specifically infringement of internet users' freedom of expression and privacy—that directly result from using the products and services of internet, mobile, and telecommunications companies. This new methodology development workstream focused on algorithms, machine learning and automated decision-making expands the scope of the RDR Index by considering human rights harms beyond freedom of expression and privacy.

Beyond UDHR Articles 12 and 19: RDR focuses on freedom of expression and privacy because the protection of these rights ensures the ability to exercise many other rights. If people's expression and privacy rights are not protected and respected, they cannot use technology effectively to exercise and defend other political, religious, economic, and social rights. Indeed, as we consider how to expand the methodology to address harms associated with the use of algorithms and related technologies, in many of the risk scenarios below, companies' failure to respect privacy and/or freedom of expression and information causes or contributes to the violation of other human rights, specifically: right to life, liberty and security of person (UDHR art. 3); non-discrimination (UDHR art. 7, art. 23); freedom of thought (UDHR art. 18); freedom of association (UDHR art. 20); right to take part in the government of one's country, directly or through freely chosen representatives (UDHR art. 21).

Types of human rights harms

The scenarios are organized into four categories, some of which overlap and intersect with one another. There is also overlap with the human risk scenarios associated with targeted advertising published in February 2019, as some uses of algorithms, machine learning and automated decision-making relate to the targeted advertising business model. In each case, company practice has a negative impact on freedom of expression and information and/or on privacy. In many cases this initial violation further affects other human rights. For each scenario, the document specifies which rights are affected and points to the relevant article in the UDHR.

A. Algorithmic curation, recommendation and ranking: Algorithmic content curation, recommendation and ranking systems that are optimized for user engagement have the effect of prioritizing controversial and inflammatory content, including content that is not protected under international human rights law. Actors who are outside of the company's control can take advantage of these systems to perpetrate or encourage human rights abuses. Over time, reliance on curation and recommendation algorithms that are optimized for engagement can alter the news and information ecosystems of entire countries or communities.

B. Algorithmic content moderation and other content restrictions: As the volume and diversity of user-generated content hosted on internet platforms grows, companies face increasing pressure from governments as well as their own business models to automate content moderation decisions through the use of algorithms, machine learning, and related artificial intelligence technologies. However, artificial intelligence (including algorithms and machine learning) is not an adequate substitute for human content moderation, leading to both false positives and false negatives. Overly broad content moderation can result in the removal or de-prioritization of protected expression, while the spread of harmful content can have severe consequences for human rights. Telecommunications companies can use algorithms, machine learning and other forms of automation to implement technical restrictions on access to online content, for example by automating the creation of “block lists” in response to government demands.

C. Violations of the purpose limitation principle: The purpose limitation principle holds that data should only be collected for specified, explicit and legitimate purposes, and that data should not be further processed in a way that is incompatible with those purposes. Companies often rely on user information in order to build and train algorithms, machine learning models, and other kinds of automated decision-making systems, and do not always obtain meaningful consent before collecting this information, much less before processing it. Moreover, user information that was collected for a specific purpose is often used for secondary purposes (such as inclusion in training datasets) to which the individual did not consent. This is a violation of the right to privacy (UDHR art. 12) that enables further harms related to manipulation, bias, and discrimination.

D. Automated manipulation, bias, and discrimination: While many people believe that algorithms are neutral decision-making agents that apply rules in an even-handed manner, they are more accurately described as “opinions embedded in mathematics.”[1] Algorithms, machine learning systems, and other forms of automated decision-making tend to replicate the logics, patterns and biases of their human designers and of the datasets on which they are trained, thus reinforcing existing systems of discrimination and oppression. Moreover, an algorithm or machine learning model that was developed for a specific context, and performs as intended within that context, can perform poorly when used in a different context and violate human rights as a result. This is true even when there is no discriminatory intent. Such “technological redlining” represents “a form of digital data discrimination, which uses our digital identities and activities to bolster inequality and oppression”.[2] Moreover, access to user information enables companies and advertisers to segment audiences in a very granular manner, tailoring messages to very specific attributes including preferences, habits, or traits (real or inferred). This in turn enables large-scale media manipulation, deliberate discrimination against internet users on the basis of protected traits, and even the targeting of specific individuals.

Risk scenarios

A. Algorithmic curation, recommendation and ranking

Scenario A1: A social media platform’s content curation and recommendation algorithm is optimized for engagement. As part of an ethnic cleansing campaign, elites from Ethnic Group A share disinformation and hate speech on the platform about the minority Ethnic Group B. This content receives high engagement, both positive and negative, and becomes “viral” as a result. False information and incitement to violence against Group B’s members spread across the platform, inspiring acts of violence against Ethnic Group B.

References:

- What happens when the government uses Facebook as a weapon?
- How Duterte used Facebook to fuel the Philippine drug war
- Soldiers in Facebook’s war on fake news are feeling overrun
- A genocide incited on Facebook, with posts from Myanmar’s military

Human rights risks: Freedom of expression and information (UDHR art. 19), right to life, liberty and security of person (UDHR art. 3), non-discrimination (UDHR art. 7, art. 23).

[1] O’Neil,Cathy. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown.

[2] Definition provided by Dr. Safiya Noble, cited in Caplan, R., Donovan, J., Hanson, L., & Matthews, J. (2018). Algorithmic accountability: A primer. *Data & Society*, p. 3. Available from https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf.

Types of companies/services: Advertising networks, services that display ads.

Scenario A2: Users seeking to learn more about their country's history visit a video-sharing platform. The recommendation algorithm suggests a series of videos that display marks of high engagement (views, comments, etc.) but are inaccurate or misleading and promote hateful views of a religious minority group. Some of these users then go on to adopt and express hateful views against members of the religious minority in question, to discriminate against them in the workplace and in daily life, or even to engage in physical violence against them.

References:

- This is exactly how social media algorithms work today
- The algorithmic rise of the "alt-right"
- Alternative Influence: Broadcasting the Reactionary Right on YouTube.
- YouTube's Alex Jones problem
- Facebook's failure to enforce its own rules
- Facebook posts "substantively contributed" to Myanmar genocide, UN investigators say
- Where countries are tinderboxes, and Facebook is a match

Human rights risks: Freedom of expression and information (UDHR art. 19), right to life, liberty and security of person (UDHR art. 3), non-discrimination (UDHR art. 7, art. 23), freedom of thought (UDHR art. 18).

Types of companies/services: Social networking sites, services that host user-generated content.

Scenario A3: A company that hosts user generated content derives much of its income from targeted advertising, the placement of which is automated. The company uses an algorithm to generate "affinity groups" that advertisers can use to target specific audiences. The algorithm recognizes certain patterns in user profiles and behavior, and determines that people who express hate toward Ethnic Group A are a valuable audience for advertisers. A hate group associated with Ethnic Group B uses this affinity group to spread hate speech targeting Group A to the users who are most likely to engage with that content. Some of these users then go on to adopt and express hateful views against members of Group A, to discriminate against them in the workplace and in daily life, or even to engage in physical violence against them

References:

- Facebook Enabled Advertisers to Reach ‘Jew Haters’
- How the online business model encourages prejudice
- What we learned from collecting 100,000 targeted Facebook ads
- Facebook will remove 5,000 ad targeting categories to prevent discrimination
- Inside the Secret Border Patrol Facebook Group Where Agents Joke About Migrant Deaths and Post Sexist Memes

Human rights risks: Privacy (UDHR art. 12), non-discrimination (UDHR art. 7, art. 23), right to life, liberty and security of person (UDHR art. 3), freedom of thought (UDHR art. 18).

Types of companies/services: Advertising networks, services that display targeted ads.

Scenario A4: A social media platform decides to scale up its algorithmically generated group recommendations in order to increase engagement among its users. The algorithms it applies operate opaquely and the company cannot predict recommendations in advance. Since social media algorithms are driven by engagement, the company’s group recommendation algorithm promotes the content that is most likely to be clicked on, without examining its impact on human rights. The group recommendation algorithm starts suggesting that users following anti-immigration and racist accounts join groups dedicated to discussing xenophobic and anti-immigrant views. As users act on the suggestion, the algorithm suggests the group to more and more users, including many who are statistically similar to group members but don’t follow any anti-immigration or racist accounts themselves. Over time, some users adopt xenophobic and anti-immigrant views, and even commit acts of violence against immigrants in their communities.

Reference(s): Inside the Secret Border Patrol Facebook Group Where Agents Joke About Migrant Deaths and Post Sexist Memes

Human rights risk: Freedom of expression and information (UDHR art. 19), non-discrimination (UDHR art. 7, art. 23), freedom of thought (UDHR art. 18).

Types of companies/services: social media platforms, messaging services offering group chat

Scenario A5: Citizens belonging to a particular minority or a marginalized group have taken to the streets to protest new policies that would threaten their human rights. A journalist notices that on platform A, which displays user generated content in a strict reverse-chronological order, footage of the protests and the police's violent crackdown were prevalent. However, on platform B, whose content curation algorithm surfaces content according to an opaque logic, the story of the protests was buried. As a result, the protesters' ability to spread their messages and reach new audiences on platform B is restricted by the platform's algorithmic filtering. Such filtering also impacts the right of users to access information about the protests.

References:

- The real bias built in at Facebook.
- What Happens to Ferguson Affects Ferguson.

Human rights risks: Freedom of expression and information (UDHR art. 19), freedom of association (UDHR art. 20).

Types of companies/services: social media platforms

B. Algorithmic content moderation and other content restrictions

Scenario B1: A company that hosts user generated content as well as advertising bans ''adult content,'' "terrorist content" and "extremist content," among other types of expression that it considers harmful. However, its automated content moderation tool persistently flags and restricts access to content that falls outside the company's stated definition of these terms. Artwork, images of women breastfeeding, content related to sexual and reproductive health, and LGBT content is incorrectly classified as "adult content," and videos or photos that document evidence of war crimes and other human rights violations are incorrectly determinted to promote terrorism or extremism. As a result, such content is consistently taken down or restricted from the platform even though it complies with the company's terms of service.

References:

- Tumblr's porn detecting AI has one job and it's bad at it.
- Facebook banned a user from posting a photo of a 30,000-year-old statue of a naked woman — and people are furious
- How Instagram May Be Unwittingly Censoring the Queer Community
- Youtube and Facebook are removing evidence of atrocities, jeopardizing cases against war criminals.
- YouTube admits 'wrong call' over deletion of Syrian war crime videos
- Google bans Irish abortion referendum adverts

Human rights risks: Freedom of expression and information (UDHR art. 19).

Types of companies/services: social media platforms, video-sharing platforms, e-commerce platforms, cloud services.

Scenario B2: Platforms A, B, C and others are part of a consortium of companies that host user generated content and cooperate to fight ‘’terrorist’’ and ‘’extremist’’ content online. The consortium creates an industry-wide hash database of content that has been identified as promoting or celebrating terrorism, allowing participating companies to create "digital fingerprints" for such content and to filter offending content prior to upload. As a result, any errors made in labeling content as ‘’terrorist’’ in this database is propagated across all other participating platforms, greatly impacting users’ freedom of expression. A journalist posts a graphic video on platform A, which depicts gross human rights violations and calls for the perpetrators to be held accountable. Platform A removes her video for violating its terms of service (which ban ‘’terrorist’’ and ‘’extremist’’ content), and uploads a hash of the video to the industry-wide database. The user also attempts to post the video on platforms B and C, whose upload filters determine that the video matches a hash in the industry-wide database and prevent the user from sharing the video, even though their policies allow users to post graphic videos for journalistic purposes. None of the companies offer any mechanism for appeal or redress.

References:

- Open letter on the Terrorism Database
- European "terrorist content" proposal is dangerous for human rights globally

Human rights risk: Freedom of expression and information (UDHR art. 19).

Types of companies/services: social media platforms, video-sharing services, cloud services

---

Scenario B3: A social media company operates in multiple countries, including a country with an ongoing conflict among several of its ethnic groups. The company uses a natural language processing (NLP) tool to enforce its content moderation rules. However, the machine learning model was trained on English-language text, and is less accurate when used to moderate content in other languages, with disparate accuracy levels for the country’s official languages. The company relies on its NLP tool during the conflict to detect harmful content, notably content that incites violence, but it is more accurate for some of the country’s languages than for others. It is least accurate for the language spoken by Ethnic Group A, and content expressed in Group A's language that incites violence against Ethnic Group B is permitted to remain on the platform. As a result, members of Group A carry out acts of violence against members of Group B.

References:

- Mixed Messages? The Limits of Automated Social Media Content Analysis
- How Facebook's rise fueled chaos and confusion in Myanmar
- The deadly Facebook cocktail: Hate Speech & Misinformation in Sri Lanka

Human rights risks: Freedom of expression and information (UDHR art. 19), right to life, liberty and security of person (UDHR art. 3), non-discrimination (UDHR art. 7, art. 23).

Types of companies/services: Social networking sites, services that host user-generated content.

C. Violations of the purpose limitation principle

Scenario C1: A company trains its algorithms through access to and analysis of previously collected user information as well as other datasets culled from publicly available information. The company's users were not informed that their information would be used in this way, nor did the data subjects whose publicly available information comprises the other datasets.

References:

- Data & Society (2018), *Algorithmic accountability: A Primer.*
- Privacy International and Article 19 (2018), *Privacy and freedom of expression in the age of artificial intellligence*
- Transgender YouTubers had their videos grabbed to train facial recognition software
- Steven M. Bellovin, et. al, "When enough is enough: Location tracking, mosaic theory, and machine learning," NYU Journal of Law and Liberty, 8(2) (2014) 555--628

Human rights risk: Privacy (UDHR art. 12).

Types of companies/services: social media platforms, messaging apps, email services, search engines, mobile ecosystems, prepaid and postpaid mobile broadband

Scenario C2: A company uses algorithms and machine learning to analyze large sets of user data to predict individual users' gender, income, location, sexual orientation, political preference, and willingness to pay for products and services. The company previously obtained meaningful and informed consent from its users to collect these types of user information, but not to include the user information in training datasets for machine learning. The company's privacy policy states that it does not process personal data because key identifying information has been removed from each record. However, the algorithmically generated predictions about the users reveal sensitive personal information, and the company is able to re-identify its users by combining different datasets. As a result, the company violates its users' right to privacy since users are not informed about the existence and operating mechanism of the algorithmic behavior prediction technology, nor are they given sufficient control over their privacy.

Reference(s):

- Facebook Uses Artificial Intelligence to Predict Your Future Actions for advertisers, says confidential document
- Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms

Human rights risks: Privacy (UDHR art. 12), non-discrimination (UDHR art. 7, art. 23).

Types of companies/services: social media platforms, messaging apps, email services, search engines, mobile ecosystems

Scenario C3: A telecommunications company and/or internet service provider uses the data it has collected about its users' communication patterns to infer their interests, without seeking or obtaining the users' consent. It then uses this information to develop new offerings that it hopes will be more profitable, to identify which users should be shown targeted advertising for these services as well as third-party targeted advertisements, and sells this inferred information about users' interests to a third party.

References:

- AT&T to end targeted ads program, give all users lowest available price

Human rights risk: Freedom of expression and information (UDHR art. 19).

Types of companies/services: telecommunications companies, internet service providers

D. Automated manipulation, bias and discrimination

Scenario D1: A company develops a computer vision algorithm to label images with text that can be used by assistive devices for vision-impaired users, among other use cases. However, the algorithm persistently mislabels images of dark-skinned people, including in ways that are demeaning or offensive.

References:

- Google apologises for Photos app's racist blunder
- Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech
- Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots
- How white engineers built racist code – and why it's dangerous for black people

Human rights risks: Privacy (UDHR art. 12), non-discrimination (UDHR art. 7, art. 23).

Types of companies/services: social media platforms, services that host user-generated content.

Scenario D2: A search engine features an auto-complete function that suggests search terms based on the first few letters that the user enters into the query. These suggestions are algorithmically generated, and the algorithm itself is a black box: there is no public visibility into the factors that contribute to each suggestion, nor can the company predict what predictions the algorithm will surface. The auto-complete algorithm frequently suggests search terms that reinforce existing social biases and structural oppression.

References:

- Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (2018). New York University Press.
- Google, democracy, and the truth about internet search

Human rights risks: Privacy (UDHR art. 12), non-discrimination (UDHR art. 7, art. 23).

Types of companies/services: social media platforms, services that host user generated content.

Scenario D3: A group of internet users who share a white male supremacist ideology develops a network of automated software agents (a "botnet") to harass and intimidate women and people of color, with the goal of bullying them into leaving a social media platform and abandoning their profession or hobby. The botnet uses machine learning, including natural language processing, to craft personalized insults based on information collected from the targeted users' online presence, and to amplify each menacing message to give the appearance of a large, angry mob threatening them.

References:

- Zoë Quinn, *Crash Overdrive* (2017). Public Affairs.
- Weaponizing the haters: The Last Jedi and the strategic politicization of pop culture through social media manipulation.
- How 2 racist trolls got a ridiculous Star Wars boycott trending on Twitter

Human rights risks: Privacy (UDHR art. 12), freedom of expression and information (UDHR art. 19), non-discrimination (UDHR art. 7, art. 23).

Types of companies/services: social media platforms, services that host user generated content.

Scenario D4: During an election campaign, a political party deploys a network of automated software agents (a "botnet") on a social networking site to boost messages that support its candidates and its policy positions, and to send demeaning and threatening responses to messages that support opposing parties, candidates or viewpoints, with the goal of creating the perception that its own views enjoy widespread support and that other views are unpopular. Some popular media outlets amplify this false narrative, reporting on the perceived trend which eventually becomes a self-fulfilling prophecy. The political party behind the botnet wins the election.

References:

- The Computational Propaganda Project
- Automating Power: Social bot interference in global politics
- Political bots spread misinformation during U.S. campaign
- The bots that are changing politics

Human rights risks: Privacy (UDHR art. 12), freedom of expression and information (UDHR art. 19), freedom of thought (UDHR art. 18), right to take part in the government of one's country, directly or through freely chosen representatives (UDHR art. 21).

Types of companies/services: social networking sites.

Scenario D5: Company A places employment ads on Company B's targeted advertising platform, using algorithmically generated targeting parameters that reflect Company A's current workforce, which mostly consists of women. As a result, individuals whom Company B does not classify as women do not know about the opportunities.

References:

- Facebook Lets Advertisers Exclude Users by Race
- Facebook Is Letting Job Advertisers Target Only Men
- Facebook Promises to Bar Advertisers From Targeting Ads by Race or Ethnicity. Again.
- How Facebook knows which apps you use, and why this matters
- Friction-free racism

Human rights risks: Privacy (UDHR art. 12), non-discrimination (UDHR art. 7, art. 23).

Types of companies/services: Advertising networks, services that display targeted ads.

Scenario D6: A platform uses its deep insight into its users' traits, behaviors, habits, and other characteristics to nudge users in Country A to vote in a national election, but does not do so in Country B, in order to measure how "effective" such nudges are.

References:

- Everything we know about Facebook's secret mood manipulation experiment
- The secret experiment behind Facebook's "I Voted sticker

Human rights risk: Privacy (UDHR art. 12), freedom of expression and information (UDHR art. 19), freedom of thought (UDHR art. 18), right to take part in the government of one's country, directly or through freely chosen representatives (UDHR art. 21).

Types of companies/services: Advertising networks, services that display targeted ads.

Scenario D7: A company's targeted advertising platform allows advertisers to target individual users on the basis of their ethnicity, which is inferred by an algorithm on the basis of their user information. In the run-up to an election, voters from Ethnic Group A are shown ads with an incorrect election date, while voters from Ethnic Group B are shown ads with the correct date. As a result, members of Ethnic Group B turn out to vote at much higher rates than Group A members, and a political party that champions Ethnic

Group B’s interests wins the election. Once in power, this party enacts policies that discriminate against Group A.

References:

- Voter suppression and racial targeting: In Facebook and Twitter’s words
- We asked for examples of election misinformation. You delivered.

Human rights risks: Privacy (UDHR art. 12), freedom of expression and information (UDHR art. 19), right to take part in the government of one’s country, directly or through freely chosen representatives (UDHR art. 21).

Types of companies/services: Advertising networks, services that display targeted ads.

Scenario D8: A social media company trains a neural network to predict the sexual orientation and race of its users without obtaining their consent, using images collected from user profile. The company violates the users’ privacy for the purpose of algorithmic surveillance by collecting and analyzing sensitive personal information and applying facial recognition. This practice not only violates users’ right to privacy but also their right to equal treatment as the company’s ad delivery system differentiates among users based on sexual orientation and race and ends up showing ads for jobs in a discriminatory manner.

References:

- Yilun Wang and Michal Kosinski, “Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation from Facial Images,” Journal of Personality and Social Psychology
- Amazon’s Face Recognition Falsely Matched 28 Members of Congress With Mugshots

Human rights risks: Privacy (UDHR art. 12), non-discrimination (UDHR art. 7, art. 23).

Types of companies/services: social media platforms, messaging apps, e-commerce platforms.

Scenario D9: Company A (the advertiser) sets up housing ads on company B’s platform targeted at various affinity groups, the composition of which is determined by company B’s algorithm. Company B sells company A the ability to target advertisements to people who share certain personal attributes and/or are likely to engage with a particular ad.

Company B provides Company A with tools to determine which users, or which types of users, Company A would like to see an ad. Company B's advertising platform enables Company A to exclude certain groups from seeing the ad based on their real or inferred gender, race or ethnicity, and location. Company B then selects from among the users eligible to see an ad which users will actually see it, with the aim of prioritizing users who are most likely to engage with the ad. Company B uses machine learning and other prediction techniques to classify and group users in order to project each user's likely response to a given ad. In doing so, Company B *de facto* creates groupings defined by race, gender, and membership in various social groups. Its automated advertising mechanisms results in the exclusion of users who are impoverished or belong to minority ethnic groups, transforming users' personal data into proxies for racial classification. As a result, users whom the machine learning model has classified as low-income or as belonging to minority ethnic groups are least likely to be shown advertisements for housing.

References:

- Facebook has been charged with housing discrimination by the US government
- Amazon doesn't consider the race of its customers. Should It?

Human rights risks: Privacy (UDHR art. 12), freedom of expression and information (UDHR art. 19), non-discrimination (UDHR art. 7, art. 23).

Types of companies/services: social media platforms, messaging apps, email services, search engines, mobile ecosystems, e-commerce platforms

---

Scenario D10: In order to optimize its return on investment, an e-commerce company decides to exclude certain neighborhoods from its same-day delivery system. It uses algorithms to identify neighborhoods where the cost of offering same-day delivery exceeds the increased sales revenue that the service would generate for the company. The company's decision-making algorithm takes the following factors into account: whether the neighbourhood has a sufficient number of members who have paid for a premium account, is located near a distribution center, and whether the company has enough delivery personnel to serve the neighbourhood. While the company's main aim is to maximize profit, it ends up excluding low-income neighborhoods inhabited by minorities. Regardless of its initial motivation, the company discriminates against racial and ethnic minorities who were excluded from the same day-delivery service because the company applies automated decision-making without assessing its wider human rights impacts.

References:

- Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms

Human rights risks: Privacy (UDHR art. 12), non-discrimination (UDHR art. 7, art. 23).

Types of companies/services: e-commerce platforms

Stakeholder consultation: We welcome feedback on these consultation documents. Feedback from a wide range of experts and stakeholders is essential to our methodology development process. After receiving feedback from experts, human rights advocates, and companies on these documents and conducting further case study research, the risk scenarios, and best practices will be used to adapt the current methodology and will be tested in a pilot study.

Please send all feedback by **September 13, 2019** to methodology@rankingdigitalrights.org.