RANKING DIGITAL RIGHTS

CONSULTATION DRAFT
Best practices for business and human rights: Algorithms, machine learning, and automated decision-making

Note: This document is the third 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 last, after the following two documents: 1) *Rationale for RDR's methodology expansion to address algorithms, machine learning, and automated decision-making* and 2): *Human Rights Risk Scenarios: Algorithms, Machine Learning, and Automated Decision-Making*. It builds on the *Risk Scenarios*—short narratives linking company practices to human rights harms—to propose concrete steps that companies should take to mitigate these harms.

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

What are "best practices"?

In the context of the RDR Index methodology development process, "best practices" are normative statements ("should" statements) about what companies should do (or refrain from doing) in order to prevent or mitigate the risks identified in the *Human Rights Risk Scenarios* document. They will form the basis for *indicators* and *elements*, which are the building blocks of the RDR Index methodology. Elements must describe practices that are technically possible for a company to implement, they must be measurable by examining the company's publicly disclosed information, and there must be a way to benchmark the disclosures of different companies against one another. Here, the best practices are organized into four categories, some of which overlap and intersect with one another, as detailed below (see *Typology of Best Practices*).

Please note that this is a draft document that will be subject to an iterative process of consultation, feedback, and revision. Best practices are intended as provocations to elicit feedback from participants in the consultation. Many represent significant departures from current company practice, and we may ultimately determine that some of the best practices listed here are out of scope for RDR or would be too difficult to evaluate using publicly available information. Please note that this document does not include best practices that are already reflected in the RDR Corporate Accountability Index methodology.

Best practices

The best practices presented below are grouped into four categories, which correspond to the harm categories presented in the *Risk Scenarios* document:

1. 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. Companies should give users the option to actively opt in to algorithmic content curation rather than having it automatically occur by default.

2. Algorithmic content moderation and other content restrictions: Companies use algorithms, machine learning, and related artificial intelligence technologies to support or augment the work of human content moderators. However, natural language processing models do not perform equally well on different types of content, depending on the language or dialect used and other factors. Companies should take steps to minimize their automated content moderation tools' differential impact on their users, offer robust appeals mechanisms, and disclose key information to their users and other stakeholders.

3. Respect 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 should commit to only using training datasets that comprise data whose data subjects have provided meaningful, informed content to having their data included in datasets used for this purpose.

4. Automated manipulation, bias and discrimination: Algorithms, machine learning models and other artificial intelligence tools are not neutral arbiters: they are "opinions embedded in mathematics"[1] that 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. In addition, the architecture of social media platforms in particular enables large-scale media manipulation, deliberate discrimination against internet users on the basis of protected traits, and even the targeting of specific individuals. Companies should take proactive measures to prevent, detect, and/or remedy such algorithmic bias, "technological redlining,"[2] and manipulation.

[1] O'Neil,Cathy. 2016. Weapons of Math Destruction. Crown.

[2] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (2018). New York: New York University Press.

1. Algorithmic curation, recommendation and ranking

Companies that use algorithmic content curation, recommendation or ranking systems should:

A. Have an easy-to-find policy document that describes how the company manages human rights risks related to the system(s). This document, and the human rights impact assessment(s) described therein, should consider the broader social contexts in which algorithmic curation, recommendation and ranking are used, in addition to the technical system itself.

B. Give users the option to actively opt in to algorithmic content curation: this function should not automatically occur by default.

C. Disclose the existence of the system to their users, describe how it works and the variables that influence the algorithm, and explain how using it will affect their user experience.

D. Make as much of the dataset(s) used to train the machine learning model available to external researchers as possible without exposing personally identifiable information.

2. Algorithmic content moderation and other content restrictions

Prior to introducing new automated content moderation tools that use algorithms or machine learning, companies should:

A. Conduct a thorough human rights impact assessment (HRIA), including an algorithmic impact assessment.

  - These HRIAs should consider the broader social context in which algorithms, machine learning models and other forms of automated decision-making tools are used, as well as the power dynamics that attend such use, in addition to fixing narrowly-defined technical systems.

  - See Fairness, Accountability, and Transparency in Machine Learning, Principles for Accountable Algorithms and a Social Impact Statement for Algorithms

B. Engage with a range of stakeholders to assess potential risks to freedom of expression and information.

C. Develop natural language processing models that are diverse and that can account for variations in language, dialect, speech and online behavior across regions, communities, and so on.

D. Test algorithmic content moderation tools *in situ* with data corresponding to the full range of languages and dialects that they will be used to help moderate.

Companies that use algorithms and machine learning for content moderation should:

A. Disclose that they use algorithms, machine learning, or other forms of automated decision-making to detect and remove content that violates their rules, and specify all the types of content they moderate with the help of algorithms.

B. Explain how algorithms and machine learning are used to support or augment the work of human content moderators.

C. Have an easy-to-find policy document that describes how it manages human rights risks related to the practice.

D. Continue to engage with stakeholders on a regular basis in order to mitigate any negative impacts on users' freedom of expression and information.

E. Test content moderation algorithms on a regular basis to ensure that the tools do not systematically perform less accurately in any language or dialect, and that they do not produce systematically different outcomes for different groups of people.

F. Regularly publish data specifying the types and number of pieces of content and accounts restricted by algorithms.

    ○ The data should specify the number of pieces of content and accounts that were removed by directly by algorithms and those that were flagged for review by human moderators.

    ○ Companies should update this data on a regular basis.

G. Notify users whose content is restricted by an automated content moderation tool that an algorithm was involved in the decision.

H. Provide all users with a mechanism to appeal the restriction of their content to a human being.

I. Regularly publish data about the volume and nature of appeals of automated content moderation decisions and the actions taken in response to those appeals.

J. Disclose key information to policymakers, researchers, and their users around the algorithms and machine learning models used for content moderation, including:

    ○ The composition of the training dataset(s) contain (e.g. how regionally, linguistically, and demographically diverse the data are);

    ○ What kind of outputs models generate;

- ○ What steps the company is taking to ensure tools are not being misused or abused in unethical ways;
- ○ Data on accuracy rates for human and automated detection and removal, including the false positive, true positive, and false negative rates, as well as the precision and recall metrics.

K. Educate policymakers and users about the limitations of automated content moderation tools.

Companies that participate in a hash sharing consortium (such as the Global Internet Forum to Combat Terrorism) should:

A. Publicly disclose their participation in the consortium.

B. Provide users who upload (or attempt to upload) content that matches a signature in the hash database with information about what entity originally entered the piece of content into the database.

C. Provide users whose content is restricted, and who believe the piece of content was incorrectly included in the hash database, with a mechanism to request the removal of the content they were trying to post from the hash database.

Telecommunications companies should:

A. Disclose whether they use algorithms, machine learning, or other automated decision-making tools such as Deep Packet Inspection (DPI) to limit users' access to certain websites or types of content.

B. Disclose whether they use algorithms, machine learning, or other automated decision-making tools such as Deep Packet Inspection (DPI) to surveil and/or manipulate users' web traffic.

3. Violations of the purpose limitation principle

Companies that develop algorithms, machine learning and automated decision-making tools should:

A. Commit to only using training datasets that comprise data whose data subjects have provided meaningful, informed content to having their data included in datasets used for this purpose.

4. Automated manipulation, bias and discrimination

In order to mitigate manipulation risks, companies should:

A. Disclose clear policies governing the use of automated software agents (“bots”) on their platforms, products and services.

B. Clearly label their own bots as being bots, and notify users that any interaction with the bot may be used to further its machine learning model.

C. Require third-party bots to be easily identifiable as such.

D. Disclose how they enforce their bot policies.

E. Engage in transparency reporting around the enforcement of their bot policies.

In order to mitigate risks related to algorithmic bias and discrimination, companies should:

A. Create and implement robust diversity and inclusion policies at every level of the company, notably in teams that develop algorithms, machine learning models, or other automated decision-making tools.

B. Conduct a human rights impact assessment prior to applying a machine learning model to data that differs from the training dataset in a significant way.

   - For example, companies should conduct such assessments before using natural language processing models that were trained on an English-language corpus to analyze text in another language.

C. Commit to examining the broader context in which algorithms, machine learning models and other forms of automated decision-making tools are used, as well as the power dynamics that attend such use, in addition to fixing narrowly-defined technical systems.

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, as we work to identify clear accountability standards that will encourage these companies to meet their obligations to respect and protect human rights. 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.