Call for Feedback: Generative AI Accountability Indicators

Thank you for your interest in providing feedback for Ranking Digital Rights's consultation on Generative AI Accountability Indicators. Feedback and input from stakeholders is essential to developing a credible, rigorous, and effective methodology—and this feedback has been integral to our methodology work since RDR's inception.

The consultation will run until Sep 10, 2023.

How to provide feedback

Begin by reading the introductory sections for important context, then closely read any indicators about which you have relevant expertise. Examine them with these questions in mind:

- Is the disclosure or policy they call for likely to be helpful in addressing human rights risks?
- Would the policies they call for cause unintended consequences?
- Do some of them seem more important than others, such that they should be given more weight in each company's final score?
- Is the standard set by the indicators technically and legally achievable by an ambitious generative AI company within two years?
- Is anything missing?

To submit your feedback, email methodology@rankingdigitalrights.org. You may specify particular indicator(s) or offer general comments. Send your feedback in written form or suggest a time for a call with the RDR team. We're happy to talk to you!

Table of Contents

About this project

In the fall of 2023, Ranking Digital Rights will release the inaugural Generative AI Accountability Scorecard, a report card and roadmap for consumer-facing generative AI services to respect the human rights to privacy, non-discrimination, freedom of expression, and freedom of information. The scorecard will rely on these indicators as a basis for scoring.

As we do currently with both our Big Tech and Telco Giants Scorecards, the project hopes to spur a race to the top among companies in this rapidly developing field. Despite the potential for many positive applications of generative AI, it has been linked to a range of human rights risks including "turbocharged information manipulation," bias, non-consensual pornography, fraud, and incentives for continued privacy violations. For more information on the rationale behind this project, see the June 2023 report published along with RDR's preliminary standards on generative AI, which formed the basis for the draft indicators.

RDR runs an email list for discussion and announcements about civil society and academic projects to evaluate the policies and transparency of generative AI services. To join, send an email to methodology@rankingdigitalrights.org.

About the indicators

Scope

These draft indicators provide guidelines for companies offering consumer-facing generative AI services, such as ChatGPT by Open AI and DreamStudio by Stability.ai, to respect the human rights to privacy, non-discrimination, and freedom of expression and information.

They apply to user-facing services generating AI that creates text or static images. Similarly to the indicators used in our existing scorecards, they are designed to examine companies' disclosed policies and transparency, excluding non-public policies and journalistic or historical accounts written by third parties.

Supply chain

These draft indicators are not designed to evaluate foundation models[1], but rather services which add extensions to foundation models to provide specialized functionality. However, they

[1] For a complementary project focused on foundation models, see the Stanford Center for Research on Foundation Models' evaluation of ten leading models' compliance with transparency requirements in the draft EU AI Act.

do expect a certain level of due diligence on the part of the service provider toward the foundation model it chooses to use.

To begin with, the draft indicators differentiate between two kinds of service providers: **deployers and flagships**. Deployers are service providers who use a third-party foundation model through an application programming interface or API (a software bridge between two companies' systems) and therefore add their own extensions to its functionality. When deployers make extensions, the foundation model provider may or may not have the ability to monitor or control them. Flagships are services that use a foundation model created by that same company. Just like deployers, flagships generally make extensions to the foundation model to specialize it for their particular service.

Scoring

The indicators are each made of questions known as "elements." On any given element, a company will receive Full credit, Partial Credit, or No Disclosure Found. Element scores will be averaged to make indicator scores, which will then be combined with a subset of RDR's existing 58 core indicators, to produce a final score for each service.

Issues not covered by the indicators

The following are a list of some of the areas about which we believe generative AI services should create particular disclosures or policies to protect digital rights. The areas covered by the draft generative AI indicators are in bold. Bolded issues are addressed in the draft indicators. Unbolded areas are not, because we believe them to already be sufficiently addressed by RDR's pre-existing core set of 58 indicators:

Model Category

- **Development transparency**
- **Bias**
- **Algorithmic auditing**
- **Informed users**
- **Consent from data subjects for training**
- **Machine Unlearning**
- **Business model**
- **Transparency for all audiences**

Accountability Category

- Human rights governance and management oversight
- **Human rights impact assessment**
- **Remedy for harms**
- **Stakeholder engagement**
- **User flagging**

Policy enforcement category

- **Policies and enforcement**
- **Account restriction appeals**
- **Data about content restrictions for policy violations**
- **Data about account restrictions for policy violations**
- Advertising policy rules and enforcement
- Process for responding to third-party requests to restrict content or accounts
- Data about third-party requests to restrict content or accounts

User privacy category

- Collection and retention of user information
- Purpose for collection of user information
- Sharing of user information
- Users' control of their own information
- Government access to user information
- Data about government access to user information

Security category

- Industry standard security
- **Addressing security vulnerabilities**
- Data breaches

Draft indicators

Definitions

- **The service:** The consumer generative AI service provided by the company.
- **The company:** Within this document, this term always refers to the company which provides the consumer generative AI service.
- **Foundation model:** The AI model which powers the service. The foundation model provider may be the company itself, or a third party whose foundation model is connected to the service through a software bridge; such as an Application Programming Interface (API).
- **Extensions of foundation models:** These are modifications done by the company to specialize the foundation model for the service. These include fine-tuning and prompt engineering; among many other methods. Extension generally makes the model perform better on a specified task which is relevant to the service, while limiting its performance in others that are irrelevant in its operation context. The foundation model provider may or may not participate in them, or even have the ability to monitor them, depending on the method of extension and the particular technology.
- **Deployers only:** Elements with this tag apply only to services which use a foundation model provided by a third-party foundation model provider, as opposed to flagship services, which use a foundation model developed by the same company that controls the service.
- **Content restrictions:** Individual actions taken by a service, whether by a human or an automated process, to prevent content from appearing on the service. Most often, this is an algorithmic rejection of a user prompt.
- **Account restrictions:** Individual actions taken by a service, whether by a human or an automated process, to limit or terminate a users' access to a service.

Model Category

M1: Development transparency

Elements:

1. Does the company disclose the foundation model its service uses, or state that it does not use a foundation model?

2. Does the company disclose what, if any, extensions it has made to the foundation model?

3. Does the company provide access to a description of the workflow that was used to train the foundation model, and of any machine learning assets that were used?

M2: Bias

Elements:

1. Does the company clearly disclose the biases of its model?

    ➢ Scoring Details: For full credit, the company should indicate that the listing is comprehensive, at least for biases above a threshold of severity, which has been clearly identified based on human rights risk. The company does not need to explain whether the biases result from the foundation model or from its own extensions.

2. Does the company explain which steps it took to identify and mitigate bias while performing extensions to the foundation model?

    ➢ Scoring Details: Partial credit should be assigned for companies explaining only identification or mitigation, rather than both. The company can get full credit for this even if it gets no credit for Element 1.

3. Does the company provide access to an explanation of which steps were taken to identify and mitigate bias while training the foundation model?

4. (Deployers only) Does the company explain its due diligence to determine that sufficient measures were taken to mitigate bias in the foundation model?

M3: Algorithmic auditing

Elements:

1. Does the company commit to regular internal assessments of the model based on metrics affecting human rights including privacy, freedom of expression, freedom of information, and freedom from discrimination as well as risks including bias, security, and the effectiveness of automated policy enforcement?

    ➢ Scoring Details: If at least one of these specific areas is addressed in the company's internal assessments, it can receive partial credit. It must address all of them to receive full credit.

2. Does the company publish results of internal assessments conducted?

    ➢ Scoring Details: The company does not need to publish the entire results of the internal assessments for full credit.

3. Does the company undergo independent third-party auditing of its model based on metrics affecting human rights including privacy, freedom of expression, freedom of information, and freedom from discrimination as well as risks including bias, security, and the effectiveness of automated policy enforcement?

- ➢ Scoring Details: If at least one of these specific areas is addressed in the company's audit, it can receive partial credit. It must address all of them to receive full credit. As long as it addresses the relevant issues, third party redteaming can earn credit as a form of auditing.

4. Does the company publish results of third-party audits, identifying areas of concern?

    - ➢ Scoring Details: The company does not need to publish the entire results of the internal assessments for full credit.

5. Does the company provide access to information about internal assessments and third-party audits undertaken by the foundation model provider?

6. (Deployers only) Does the company disclose its due diligence to determine that the internal assessments or third-party audits carried out by the foundation model provider were adequate?

M4: Informed users

Elements:

1. Is it clear, from within the user interface of the service, that the user is interacting with a machine?

    - ➢ Scoring Details: If the interface for the service is publicly accessible without requiring that the user make an account, this can be evaluated directly by researchers. Otherwise, for full credit, the company must explain, on a publicly accessible web page, how it ensures users know they are interacting with a machine, such as with a screenshot of the interface. Services where the user is obviously interacting with a machine, such as those with "AI" or "bot" in their name, should get full credit automatically.

2. Does the company explain how accurately the outputs of the service are likely to reflect reality?

3. Does the service explain whether the service is likely to produce different quality of output for certain types of input?

    - ➢ Scoring Details: For example, "This system is best at generating pictures of animals, but doesn't do well at making pictures of people."

M5: Consent from data subjects for training

Elements:

1. Does the company explain which personal information was used to extend the foundation model?

2. Does the company explain how it obtained affirmative consent from people whose information it used to create any extensions to the foundation model?

    ➢ Scoring details: Affirmative consent means that a person took an action to say they consented, rather than choosing not to take an action. For example, checking a box rather than letting a pre-checked box stay checked.

3. Does the company provide access to an explanation of whether and how the foundation model provider sought consent from people whose information was used to train the foundation model?

    ➢ Scoring Details: For full credit, the company's foundation model provider does not need to say that it obtained affirmative consent from all people whose data was used in training. It only needs to clearly describe a reasonable and non-deceptive approach to consent. In the typical case of a foundation model trained on data scraped from the public internet without consent, a company can earn full credit as long as its foundation model clearly explains why it was not able to seek consent.

4. If the company does not provide evidence that the foundation model provider sought affirmative, opt-in consent for all people whose information was used to train the foundation model, does it provide access to a justification of this decision?

    ➢ Scoring Details: If the company receives full credit on Element 3, this element is not applicable.

Methodology development notes

Historically, RDR has called for companies not to use personal information to train machine learning algorithms without explicit, opt-in consent (see Indicator P7 from the RDR Corporate Accountability Index methodology). This is a higher bar than the European Union's General Data Protection Regulation (GDPR), the world's most prominent data protection law, which offers legal bases for using data other than consent. We maintained our standard in Element 2, which applies to the consent sought by the company for the use of personal information while extending the foundation model.

However, because of the huge amount of publicly visible, web-scraped data used to train most foundation models, this standard would likely be impossible for most foundation

model providers to meet at the present time and into the near future. Because it is important that RDR standards be attainable, we lowered the standard for foundation model providers in Elements 3 and 4. Companies can get full credit if they are able to provide access to a statement from the foundation model provider they use, explaining that the provider was not able to get consent due to the scale of the training dataset, and providing a justification for training the model anyways.

M6: Machine unlearning

Elements:

1. Does the company explain how individuals can promptly remove the influence of information about them on any extensions it made to the foundation model?

    - ➢ Scoring Details: If this process is only accessible to users, the company should get partial credit.

2. Does the company disclose any limited conditions, in which it denies requests from individuals to remove the influence of information about them on any extensions it made to the foundation model?

3. Does the company provide access to a method for users to remove the influence of information about them on the foundation model?

    - ➢ Scoring Details: The company can still get full credit if the foundation model specifies a significant time delay or has exceptions.

Scoring Details: Elements 1 and 2 refer to the personal information used by the company during training or other operations performed when making extensions to the foundation model, while Element 3 refers only to personal information used in the training of the foundation model.

Methodology development notes

Similarly to Indicator M5 on consent, this indicator seeks a balance between the current state of generative AI and the ideal scenario, as determined by our interpretation of human rights principles.

The EU's General Data Protection Regulation (GDPR) is the most prominent piece of legislation that covers the right to machine unlearning. However, this coverage relies on an interpretation of the GDPR's right to erasure rather than being explicitly set forth in the statute. The right to machine unlearning has not yet been enforced since the GDPR came into effect in 2018, and AI developers have been able to continue using models without providing this right. In the interest of protecting human rights in this rapidly evolving field, this indicator expects companies to provide machine unlearning, but

Element 3 recognizes that, in limited circumstances, this may be impossible, and allows the company to earn some credit for explaining those circumstances.

Additional training of existing foundation models, to extend them for a specific purpose, requires much less data than the original training of the foundation model. This means that Element 3, which deals with extensions to foundation models, is easier for companies to earn credit on than Element 4 concerning foundation models themselves. We believe that, in the near term, the only feasible way for prominent foundation models to offer machine unlearning capabilities is to pledge to remove personal information from the next iteration of the model, rather than retraining on demand. Because of that, research guidance for Element 4 allows companies to get full credit if the foundation model provider they use bestows machine unlearning with a time delay.

M7: Business model transparency

Elements:

1. Does the company disclose how the service makes money for the company?
2. Does the company disclose whether the model is designed to encourage particular user behaviors?
    - ➢ Scoring Details: for example, clicking a link or divulging certain information.

M8: Transparency for all audiences

Elements:

1. Does the company provide access to non-technical, accessible language describing the service and how it functions?
2. Does the company provide access to a description of the service and how it functions, tailored to an expert audience?

Accountability Category

A1: Human rights impact assessment

Elements:

1. Does the company assess freedom of expression and information risks associated with the service?
2. Does the company assess privacy risks associated with the service?
3. Does the company assess discrimination risks associated with the service?
4. Does the company conduct assessments on a regular basis?
5. Does the company provide access to an explanation of how the foundation model provider performs impact assessments on the foundation model?
6. (Deployers only) Does the company explain its due diligence to ensure that the foundation model provider's human rights impact assessment processes are adequate?

A2: Remedy for harms

Scoring Details: the remedies covered in this indicator do not include appeals for content or account restrictions, which are covered in indicator PE2.

Elements:

1. Does the company disclose it has a grievance mechanism(s) enabling users to submit complaints if they feel their human rights, including privacy, freedom of expression, freedom of information, and non-discrimination, have been adversely affected by the company's policies or practices?
    - ➢ Scoring Details: To receive full credit, companies do not have to address the full list of rights mentioned, but should address human rights or, at minimum, some of the specific rights listed. For example, if the company says it will protect human rights through grievance mechanisms, a commitment to a certain right plus an open-ended grievance mechanism should be enough for full credit.
2. Does the company disclose its procedures for providing remedy for rights-related grievances?
    - ➢ Scoring Details: A robust appeals process should include oversight by a human reviewer and give affected users an opportunity to present additional information.
3. Does the company disclose timeframes for its grievance and remedy procedures?

- ➢ Scoring Details: Companies should also offer a clear timeframe for reviewing appeals and clearly disclose the circumstances in which appeals are not possible.

4. Does the company clearly disclose the number of complaints received related to rights violations?
5. Does the company disclose evidence that it is providing remedy for grievances?
6. (Deployers only) Does the company commit to managing any grievance complaints on behalf of the user that require interacting with the foundation model provider?

Methodology development notes

Element 6 is included to clarify that the ultimate responsibility for users' experience in the service lies with the service provider, not the provider of the foundation model or any other tools used by the service.

A3: Stakeholder engagement

Elements:

1. Does the company disclose that it has engaged with civil society or stakeholders that represent, advocate on behalf of, or are people whose human rights are directly impacted by the company's business?
    - ➢ Scoring Details: Examples of this type of engagement would be ad-hoc participation in civil society initiatives or bilateral meetings with stakeholders. For full credit, it must be clear that the engagement involves issues of privacy, freedom of expression, non-discrimination, or freedom of information.
2. If the company engages with civil society or stakeholders, is that engagement systematic and sustained?
    - ➢ Scoring Details: An example of systematic and sustained engagement would be participating in a multistakeholder forum with civil society, academia, government, and other companies.

Methodology development notes

At the time of writing, no dedicated forum exists for generative AI companies to engage with multiple types of stakeholders (civil society, academia, governments, and other companies), but we hope that this indicator encourages companies to help start one.

A4: User flagging

Elements:

1. Does the company provide access to a mechanism for users to report problematic behavior by the model?
2. Does the company disclose the process and timeframe for reviewing reports made by users?
3. Does the company disclose how complaints are reviewed, including the role of automation in the process?
4. Does the company disclose that it will notify users of the decision made regarding their report?

Methodology development notes

We opted against expecting companies to provide a way to report problematic behavior by other users of the service, because it appears that this could only happen through features that were not core to the generative AI service, such as chat rooms attached to the service. Only a limited number of generative AI services have such spaces, and RDR's core 58 indicators would suffice to address harms perpetrated by one user against another within those spaces.

Policy enforcement category

PE1: Policies and enforcement

Scoring Details: This does not apply to any restrictions imposed on advertisements or advertising accounts which operate on the service.

Elements:

1. Does the company disclose what user activities are not allowed on the service?
2. Does the company comprehensively disclose the range of content restriction methods it uses to enforce its policies?
    - ➢ Scoring Details: These include algorithmic responses to user activity, such as refusing to respond to a prompt.
3. Does the company comprehensively disclose the range of account restriction methods it uses to enforce its policies?
    - ➢ Scoring Details: For example, banning accounts or temporarily suspending accounts.
4. (Deployers only) Does the company provide access to an explanation of any policies and enforcement measures that take place at the foundation model level?
5. (Deployers only) Does the company explain how the foundation model's policies work in tandem to its own policies and enforcement processes?

Methodology development notes

Early versions of this indicator included some expectation that the service explain when it might proactively report users' activity to government agencies without a request. The current version omits for clarity, allowing this indicator to focus only on the company's internal enforcement of its own policies.

PE2: Account restrictions appeals

Elements:

1. Does the company clearly disclose that it offers affected users the ability to appeal account restrictions?
2. Does the company clearly disclose circumstances when appeals to account restrictions are not permitted?
3. Does the company clearly disclose its process for reviewing appeals to account restrictions?
4. Does the company clearly disclose its timeframe for reviewing appeals to account restrictions?
5. Does the company clearly disclose that it provides the affected users with a statement outlining the reason for its decision, when it restricts their accounts?
6. Does the company publish data on the number of appeals it received, with the percentage of appeals that it granted?

Methodology development notes

Early versions of this methodology included a sister indicator to this one, which evaluated the provision of appeals for content restrictions. We later decided that, while content policy is important in the generative AI context, the vast majority of individual content restriction actions are unlikely to have a significant impact on human rights. Users' freedom of expression and information is much more likely to be damaged by a mistaken decision to suspend their account than a denial to generate a response to a specific prompt the system mistakenly believes to violate its rules. We dropped the sister indicator in the interest of keeping the methodology focused on the most important issues.

PE3: Data about content restrictions for policy violations

Elements:

1. For each content restriction method the company uses, does it publish data about the total number of times it took that action?

    - ➢ Scoring Details: For full credit, the disclosure should include numbers for each content restriction method identified in PE1.2.

2. Does the company publish data on the number of content restriction actions it took, based on which rule was violated?

3. (Deployers only) Does the company provide access to data about content restricted by the foundation model for violating its own rules?

    Scoring Details: If it is unclear from the company's disclosures whether the data refers to content restrictions carried out by the service or foundation model, such disclosures should be considered by Element 2.

PE4: Data about account restrictions for policy violations

Elements:

1. For each account restriction method the company uses, does it publish data about the total number of times it took that action?

    a. Scoring Details: For full credit, the disclosure should include numbers for each account restriction method identified in PE1.3.

2. Does the company publish data on the number of account restriction actions it took based on which rule was violated?

Methodology development notes

This indicator does not include any elements related to the foundation model because it is unlikely that a foundational model could restrict the user accounts of a deploying service.

Security category

S2: Addressing security vulnerabilities

Elements:

1. Does the company clearly disclose that it has a mechanism through which security researchers can submit vulnerabilities they discover?
2. Does the company clearly disclose the timeframe in which it will review reports of vulnerabilities?
3. Does the company commit not to pursue legal action against researchers who report vulnerabilities within the terms of the company's reporting mechanism?
4. (Deployers only) Does the company disclose doing due diligence on the practices of the foundation model provider with regard to vulnerability reporting and justify its decision to use the foundation model provider?

Methodology development notes

The indicator expects companies not to pursue legal action against people who submit good faith reports of vulnerabilities. We considered expecting the same level of protection for researchers who report vulnerabilities that ultimately stem from the foundation model. That would mean that, to get full credit on this indicator, a deployer would need to opt not to use foundation model providers that did not meet this standard. We determined this was too far outside of the bounds of the responsibilities of deployers, and softened the expectation so that they can get full credit for performing due diligence on the practices of the foundation model provider with regard to vulnerability reporting.

About Ranking Digital Rights

Ranking Digital Rights fights for corporate accountability in the digital age. Our primary products are our Big Tech Scorecard and Telco Giants Scorecard, which evaluate 26 of the world's biggest companies on their respect for the human rights to privacy and freedom of expression. Each is released every two years, as of 2024, using hundreds of metrics to examine companies' transparency and disclosed policies. The Generative AI Accountability Indicators are intended as an independent supplement to the core indicators used in the Big Tech and Telco Giants scorecards.