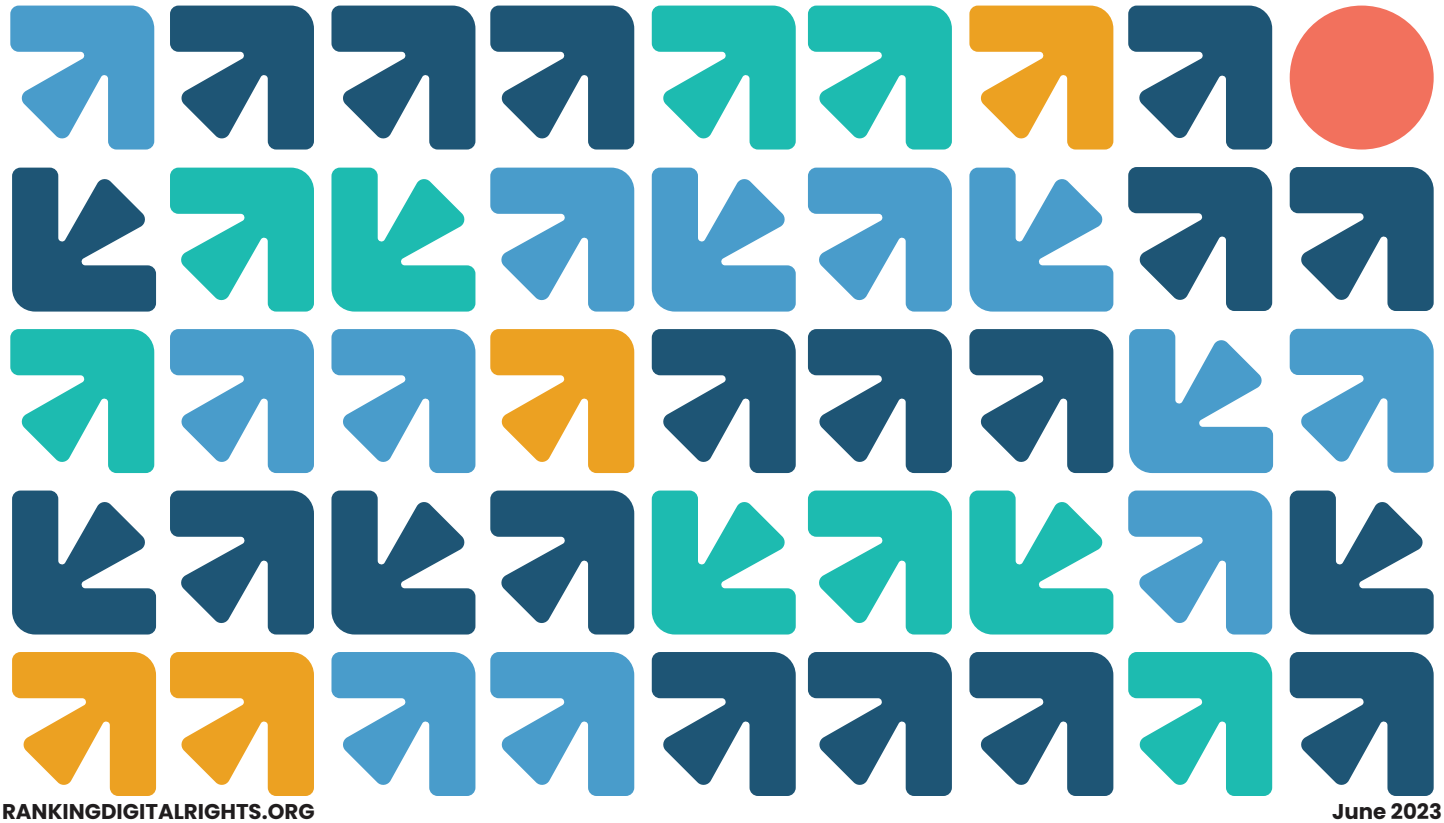
Ranking Digital Rights

# Preliminary Standards for Generative AI



# Preliminary Standards for Generative AI

## About the standards

These preliminary standards provide guidelines for companies offering consumers 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 are not part of the RDR Corporate Accountability Scorecards, but rather complementary, providing more specialized guidelines for generative AI services.

These standards apply, at the moment, only to generative AI that creates text or static images. They apply to user-facing services, not to models used internally by companies or labs.

Many consumer generative AI services make use of third-party foundation models through an Application Programming Interface. However, the services have the capacity to make significant extensions to the behavior of the original foundation model, with or without the knowledge of the company providing the foundation model. These standards are not for foundation model providers, however, they do impose obligations on service providers to perform due diligence on the foundation models they use, and cooperate with the foundation model provider in creating a transparent and rights-respecting service for users. For flagship consumer generative AI services, which are offered by the same company as the foundation model, the standards still apply, but they should be interpreted as calling for explicit disclosure of activities by the service provider, rather than due diligence to ensure that a third party foundation model provider has done them.

The standards are general statements rather than precise indicators like the ones used in the RDR Index. Similarly to the indicators used in our scorecards, they are designed to examine

companies' disclosed policies and transparency, excluding non-public policies and journalistic or historical accounts written by third-parties. In the future, following a civil society consultation, RDR plans to develop them into indicators that can be used for precise measurement and comparison. Both the preliminary standards and eventual indicators are also intended to inspire future regulation.

## Definitions

- **The company:** Within these standards, this term always refers to the company which provides the consumer generative AI service.
- **The service:** the consumer generative AI service provided by the company.
- **Foundation model:** a powerful generative AI model which powers the service. The foundation model provider may be the company or a third party whose foundation model is connected to the service through a software bridge, such as an Application Programming Interface (API).
- **"Regarding foundation models:"** Where this tag appears on subparagraphs throughout the standards, it refers to obligations of the company with respect to the foundation model it uses. These apply regardless of whether the foundation model provider and the service are owned by the same entity, though the exact form the disclosures and policies take may vary.
- **Extension 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.

# Table of Contents

# Model Category

## Development

The company should disclose which foundation model the service is powered by and clearly identify the foundation model provider (which may be the same company in the case of a flagship service). It should explain any fine-tuning or other extension it has done to the foundation model.

> *Regarding foundation models:* The company should provide easy access to a description of the workflow that the foundation model provider followed to train it, and of the datasets, pre-existing models, software packages, and tools that were used.

## Bias

The company should disclose the steps taken to identify and mitigate bias that could have been introduced during its own extension of the foundation model. It should explain any known biases that it was not able to mitigate.

> *Regarding foundation models:* The company should provide easy access to information about analogous measures taken during the creation of the foundation model, as well as a description of the company's due diligence to ensure these measures were adequate.

## Auditing

The company should commit to conducting regular internal assessment and independent third-party auditing of the model, on metrics affecting the human rights to privacy, freedom of expression, freedom of information, and freedom from discrimination. These assessments and audits should consider bias, security, and effectiveness of automated policy enforcement

(such as technical safeguards that prevent the service from outputting harmful content). It should publish summary findings of the assessments and audits, including areas of concern that were identified.

> *Regarding foundation models:* The company should provide easy access to information about analogous measures taken by the foundation model provider, as well as a description of the company's due diligence to ensure these measures were adequate.

### Informed users

The company should inform users that they are interacting with a machine, explain in detailed but clear terms the extent to which the information or depictions it provides are likely to reflect reality, and disclose any other limitations of the model.

### Consent from data subjects for training

The company should clearly disclose what personal information was included in the training data used during the extension of the foundation model. It should describe the information and whether it obtained consent from data subjects, as well as whether they were users of the system or non-users.

### Machine Unlearning

For individuals whose data was used by the company while extending a foundation model, the company should provide the ability to remove the influence of their personal information on the algorithm's behavior. The company may justify a time delay or exceptions to this capability, such as for cultural works or public figures. This right should be extended both to people who use and do not use the service.

> *Regarding foundation models:* The company should provide easy access to a method for its users to have the influence of their data removed from the foundation model. This

includes personal information included in large public datasets such as the Common Crawl. Because of the expense of retraining foundation models, broader exceptions and longer time delays may be justified.

### Business model

The company should disclose in precise language how the model makes money for the service, and whether (especially in the case of conversational generative AI services) it is programmed to steer user interactions toward particular behaviors, such as clicking on a link or divulging certain information.

### Transparency for all audiences

The company should use non-technical, accessible language to describe the model and its function, while also providing a more technical description for expert audiences. For example, a company might provide a video explaining generative AI for beginners, alongside a whitepaper explaining the extensions it made to a foundation model.

*Regarding foundation models:* The company should provide easy access to similar information for the foundation model it uses.

## Accountability Category

### Human rights governance

The company should publish a formal commitment to respect the human rights of users and non-users of its service, specifically mentioning privacy, non-discrimination, freedom of expression, and freedom of information. It should have internal structures to enforce this policy, including training and whistleblower programs. Specific positions at every level of the organization should have clear responsibility for enacting the commitment.

### Human rights impact assessment

The company should disclose that it conducts regular, comprehensive, and credible human rights impact assessments, to identify how its development and use of algorithmic systems affect human rights. The company should clearly commit to modify or terminate projects assessed as likely to harm users or non-users.

> *Regarding foundation models:* The impact assessments should consider the entire functioning of the system, including the foundation model. Since the company must rely on information supplied by the foundation model provider in this assessment, it should explain its own due diligence processes for assessing the trustworthiness of the information provided.

### Remedy for harms

The company should explain clear, prompt, and predictable grievance and remedy mechanisms to address direct harms caused by its service to users' human rights, including privacy, freedom of expression, freedom of information, and non-discrimination.

> *Regarding foundation models:* if seeking remedy requires interacting with the foundation model provider, the company should commit to handle this on behalf of the user.

### Stakeholder engagement

The company should explain how it engages with a range of stakeholders, especially civil society organizations representing people — both users and non-users — who may be adversely affected by its generative AI service. It should maintain dialogue with them

regarding the service's impact on freedom of expression, freedom of information, privacy, and non-discrimination.

### User flagging

The company should disclose that it provides users with a method, within the service's interface, to report problematic behavior by the model or other users. This system should include a way for users to track the reports and whether action was taken based on them.

## Policy enforcement category

### Policies and enforcement

The company should comprehensively disclose what user activities are not allowed on the service and all measures it takes to enforce these policies. These include algorithmic responses to user activity, such as refusing to respond to certain prompts, issuing warnings in response to certain prompts, or reporting certain user behavior to a human reviewer or to law enforcement. They also include human decisions, such as to restrict accounts.

> *Regarding foundation models:* The disclosure should include any relevant policies and enforcement which takes place at the level of the foundation model, and how it works in tandem with its own policies and enforcement.

### Appeal

The company should commit that, if it restricts a user's account in response to a policy violation, it will explain the decision to them and offer them the opportunity for a prompt appeal.

### Data about policy enforcement

The company should clearly disclose and regularly publish data about the volume and nature of actions taken to enforce its policies, and the policies of the foundation model provider, with regard to its users.

> *Regarding foundation models:* The disclosure should include and distinguish enforcement that takes place at the level of the foundation model. For example, in reporting the number of times that its service refused to respond to a users' prompt, it should separate the number of such prompts which were denied by its own software, from the number which reached the foundation model and were denied by it.

## User privacy category

The user information considered in this category is collected by the company directly from the user or from third parties. This is distinct from personal information ingested from large datasets, such as Common Crawl, often used during training of foundation models.

### Collection of user user information

The company should clearly disclose what information it collects about users, including from the prompts they provide to its generative AI, and for what purposes it collects this information. This should include a commitment to data minimization.

### Users' control of their own information

The company should allow users to obtain all of their user information it holds, control its collection and processing, and delete it.

### Government access to user information

The company should disclose what access governments have to the user information it holds, and regularly publish data about the incidence of this access. This includes proactive reporting of user behavior to governments without receiving a demand, such as to interrupt a suicide attempt revealed in prompts or to report someone trying to use the algorithm illegally.

## Security category

### Industry standard security

The company should explain how it upholds a high standard of information security, including training, third-party certifications, and audits.

### Addressing security vulnerabilities

The company should provide a method for external security researchers to submit good-faith reports of vulnerabilities in the software providing the service, free from legal threats.

> *Regarding foundation models:* The company should maintain an open channel of communication with the foundation model provider, so that vulnerabilities determined to be stemming from the foundation model can be promptly addressed.

### Data breaches

The company should disclose information about its processes for responding to data breaches. This should include reporting to governments, and efforts to inform and assist affected people.

Ranking Digital Rights



RANKINGDIGITALRIGHTS.ORG