HUGGING FACE

Hugging Face S.A.S
9 rue des colonnes,
75002 Paris, France

January 24., 2023

Hugging Face Feedback on the Digital Services Acts - Transparency Reports

About Hugging Face

Hugging Face is a community-oriented company based in the U.S. and France working to democratize good Machine Learning (ML), and has become the most widely used platform for sharing and collaborating on ML systems. We are an open-source and open-science platform hosting machine learning models and datasets within an infrastructure that supports easily processing and analyzing them; conducting novel AI research; and providing educational resources, courses, and tooling to lower the barrier for all backgrounds to contribute to AI. As part of our activity, Hugging Face provides social features and communication platforms for people interacting with AI systems, including social posts and discussion threads, in addition to hosting the AI systems themselves.

Comments on “Draft implementing regulation - Ares(2023)8428591.pdf”

On adding further content:
We suggest that transparency reports provide readers with relevant information about the recommender systems that manage the content on their platform. Consider requiring documentation in the transparency reports for any relevant parameters of recommendation systems that affect content moderation, as well as information about how the moderation action may affect future recommendations. This could be, for example, at the same level of granularity as when applying recital 70 and article 27(2).

On “machine readable” formats for reports:
The current framing explains, “to ensure that the transparency reports are machine-readable”, transparency reports are required to be in a CSV (comma-separated values) format. We agree this idea is good at a high-level, but in practice csv formats can have issues in appropriate rendering (for example if there is content with additional commas, non ASCII-characters, etc.) This also misses the option to have easily human-readable reports that enable people to interact with the content, such as by printing it out, highlighting, making notes, etc. much easier.

Our Suggestion is to support this requirement by creating a script that renders an input .csv transparency report as a human-readable pdf. In a user interaction with this script, users and organizations would use an upload portal with this script embedded to upload their .csv transparency report, which then renders it in a human-readable pdf for the organization to check. If the rendering has errors, this is on the organization's side to fix before their final submission. This removes the burden from the government's side of trying to solve for another organization's errors, and makes it much easier for the organization to find issues. This also ensures human-readability as well as machine-readability.

Comments on Annex 1 - Templates

Regarding 2_categories_names

The categories bring up several issues. Two common themes emerge across them:

1. **The distinctions between illegal and legal content.** Legal and illegal content are grouped together in this categorization. Sometimes "illegal" is mentioned specifically; sometimes it is not. We recommend keeping these two distinct. The reasons for this include:
    a. Not requiring service providers to make the decision as to whether something is technically illegal or not
    b. Not creating reporting requirements for legal and appropriate content
    c. Not creating (additional) stigma against legal activities

    Further, for services that are offered in more than one country in the EU, the definition of 'illegal' may vary, causing potential confusion as to which definition of illegal should be considered in the report.

    Our Suggestion: If there is a goal for service providers to document illegal/unflawful content explicitly, consider revising the instructions to specify that service providers should report all illegal content they are aware of for each category.

2. **Domains and stakeholders**.
    a. **Modality:** A focus here is on *image* content. Consider removing the focus solely on images, or else explicitly including additional modalities where the harms of concern are also expressed, such as audio.
    b. Stakeholders: In all categories, consider additionally grounding on EU "protected grounds" and/or "sensitive personal characteristics". This includes categories such as those listed as sensitive personal data or subject to

non-discrimination by the European Commission or in European non-discrimination law. This would also mean updating the subcategory of *"Gender-based violence"* to something like *"Violence based on protected grounds or sensitive characteristics"* and the subcategory of *"Biometric data breach"* to something like *"Protected personal data breach"*.

We also note that these categories will have overlapping issues, which may make reporting a bit more confusing using this format. For example, *"Negative effects on civic discourse or elections"* could include racist misinformation, which would also fall under *"Illegal or harmful speech"*. Similarly, if the same content is considered in multiple categories, it seems this would inflate the number of notices: One report on a comment that has race and gender based discrimination could look like 2 notices in the report.

Category-specific suggestions:

- **Animal Welfare**
  - **Unlawful sale of animals:** Consider "Unauthorized".
- **Data protection and privacy violations**
  - **Biometric data breach:** As mentioned above, consider other categories of sensitive data per the GDPR. Consider specifying also genetic, health related data.
  - **Missing processing ground for data:** Consider removing. Platform users are generally not aware of what data is collected and even less aware of how it is based on processing ground. We wonder whether this is likely to be a very low number compared to other categories, which will give the impression that data is generally collected lawfully because there are fewer reports.
  - Also consider including **doxxing** as a specific subcategory.
- **Risk for public security**
  - **Risk for environmental damage:** Consider making this its own superordinate category, with subcategories corresponding to carbon, energy, effects on air, effects on water, effects on animals and plant life, etc.
- **Pornography or sexualized content:** This category appears to refer to legal content that does not create harm. We disagree that this category fits within the context of the others. Going a step further, we disagree that the government should play a role in regulating legal expressions of sexuality on platforms that permit such content. If reporting of legal sexual content is desired, consider requiring it if it can be classified within the category "Content in violation of the platform's terms and conditions" – for example, "Pornography or sexualized content" might replace the "Nudity" subcategory there.
  - **Image-based sexual abuse:** Consider "sexual abuse" as the subcategory, removing the requirement that it must be image-based to be reported.
- **Self-harm**
  - This is a type of violence. Consider combining with the Violence category.
- **Violence**

- **Gender-based violence:** As mentioned above, consider updating to include violence based in any "protected" or “sensitive” characteristic, such as would include religion, race, etc.

Regarding 6_overall_figures and 8_by_language_and_country Templates

On “accuracy rates” and “error rates”:
In statistical analysis, what you are referring to as “accuracy rate” is more commonly called “accuracy”. Reporting accuracy and error rates can be somewhat redundant; `error rate = 1-accuracy`.
Our Suggestion is, in addition to Accuracy, consider the following metrics to be required, which should be defined based on are more sensitive to the harms of different kinds of errors and can be selected based on the application:

- sensitivity, recall, hit rate, or true positive rate (TPR)
- specificity, selectivity or true negative rate (TNR)
- precision or positive predictive value (PPV)
- negative predictive value (NPV)
- miss rate or false negative rate (FNR)
- fall-out or false positive rate (FPR)
- false discovery rate (FDR)
- false omission rate (FOR)

These should be disaggregated by various factors, such as via language and sensitive/protected demographic categories.

Comments on Annex II

Regarding Section 1.7

On “linguistic expertise”:
Annex II notes that “Article 42(2)(b) prescribes that transparency report shall include the linguistic expertise of the persons carrying out the activities referred to in point (a).”
Our Suggestion is to center on *language fluency* rather than *linguistic expertise*. The former deals with language proficiency as a function of a person’s culture, upbringing, experience, etc. The latter is a field of study concerned with syntactic and morphological structures, etc. Further, for content that is not language-based, such as images, expertise such as *cultural knowledge* may be more relevant.

On “Measures taken to provide training and assistance to persons in charge of content moderation”:
This should perhaps have a "quantitative" complement, which would include compensation and working hours.

On “Summary of the content moderation governance structure”: This is fundamental to everything else the transparency reporting is looking for; it should perhaps be more highlighted, and with more explicit instructions on how to document this.

Regarding Section 1.8

On disaggregated by language:
This is a critical point that is often overlooked, and we are glad to see it in-place.