Summary of changes

We had submitted a previous version of this paper to FAccT 2023. R1 (who seems to have a legal background) gave a high score and mentioned that our analysis is in-depth and our methodology is novel. R2 gave a reject decision, and R3 was neutral (both seemed to have computer science backgrounds). Based on the reviews from R2 and R3, we realized that we needed to make the main contributions of our paper clearer to all audiences. Like our current paper, our previous paper was also half legal analysis and half technical analysis. We have edited the paper to make our contributions and argumentation clear to both computer scientists and legal scholars. In our new submission:

1. Our legal analysis is now more succinct and has an explicit purpose (changes intended for the CS readers, which were obvious to the legal scholars). We now motivate our legal analysis with the goal of preventing deficient reporting from social media companies in fulfilling the DSA transparency requirement.While this is one of many possible purposes for this legal analysis, we believe this focus makes our work more accessible to a broader audience.
2. Why our legal analysis is concerned with legislative intent is now made explicit. R3, for instance, was confused why we needed a legal analysis at all if the DSA requirement already permits alternative metrics. To summarize, our analysis first elucidates legislative intent, which then gives legal justification to alternative metrics. Appropriate legal justification is necessary since we hope that our analysis can be incorporated into future legal guidelines or amendments.
3. Section 3 now clarifies the connection between transparency reporting and statistical estimation. R2 and R3, for instance, misunderstood why stratified sampling is needed at all. We clarify that moderation recall needs to be estimated both unbiased and efficiently. Unbiasedness is needed to ensure truthful reporting for legal requirements, and unbiased estimation needs to be efficient to respect companies' right to business.
4. Our technical results are now recast exclusively in terms of moderation recall. Previously we provided all our results in terms of prevalence, which can be used to estimate recall. Now, our experiments directly estimate recall instead of prevalence. Our paper also now has an expanded discussion on precision (for purposes of completeness).
5. We include a policy discussion in the broader perspectives. While R2 is correct that recall still falls short to capture many nuances in content moderation, we discuss why it is an improvement over the current count-based metrics companies report now. This section also includes insights for those who may be drafting new transparency requirements. As of now the metrics required in law are primitive and necessitate simplification, but we hope others can build on our work.

SUBMISSION: 121
TITLE: Operationalizing content moderation ``accuracy'' in the Digital Services Act

------------------------ METAREVIEW ------------------------
The paper examines the obligation to report on content moderation “accuracy” under the DSA. My decision is to reject the paper. I agree with the reviewers that the paper generates a very interesting conversation about what “accuracy” means from a technical perspective and how it should be implemented. However, the reviewers (with more technical expertise than myself) consider that the methodology is flawed and the technical analysis is too narrow. I hope the authors will find the reviews helpful to keep working on the paper. The kind of interdisciplinary project that the authors undertake is essential for the successful implementation of the DSA.

---------------------- REVIEW 1 ---------------------
SUBMISSION: 121
TITLE: Operationalizing content moderation ``accuracy'' in the Digital Services Act
AUTHORS: Johnny Wei, Frederike Zufall and Robin Jia

----------- Summary -----------
The paper explores how a technical implementation could be materialized for the "accuracy" requirement in the recently adopted Digital Services Act. The paper therefore provides a timely analysis that explores the practical elements of recent European legislation with a global impact. The authors clearly illustrate that accuracy is not the correct frame within the legal framework, and a shift is required towards other concepts (e.g., precision, recall). The topic and methodology of this submission is novel, especially as the Digital Services Act was adopted in 2022, and as technical explorations of European legislation are gaining relevance across multiple disciplines.
----------- Contributions -----------
(A) Identification or articulation of a novel problem, phenomenon, or potential societal harm;
(C) Novel problem identified;
(F) Datasets and/or evaluation strategies that provide novel insight;
----------- Relevance: The submission has relevance to FAccT -----------
SCORE: 3 (strongly agree)
----------- Overall evaluation Score -----------
SCORE: 2 (accept: Would be a top 50% accepted paper at venues I highly respect. A very good submission, clear accept. I vote and argue for accepting this submission.)
----------- Reviewer's confidence -----------
SCORE: 3 (medium: You are fairly confident in your assessment, though the paper is not in your specific field of expertise. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.)
----------- Review -----------

S1: The paper is well-written and provides an in-depth analysis of both the legal rules and the technical implementation of those rules. Both elements are thoroughly investigated; the authors illustrate excellent knowledge of both fields.

S2: The authors clearly set out the relevant legal rules and identify where the possible problems lie when the rules are translated to technical implementation. The authors then offer recommendations for the legal framework that are grounded in well-written and thoroughly tested operationalizations of accuracy. Such empirical work to support legal frameworks is highly relevant for modern legal framework that cross digital and technical boundaries and the authors therefore contribute to multidisciplinary scholarship.

O1: Section 2.1. The authors might want to consider EU-equivalent rules for Section 230 of the Communications Decency Act. The lack of accountability for content hosted on specific platforms also existed in the EU before the Digital Services Act and continues in the DSA. The DSA, in principle, holds that hosting providers are not liable for content... as long as they comply with the DSA's obligations (Article 6). The principle thought, therefore, aligns with CDA Section 230.
The "no regulation" frame, in turn, needs revision. First, the DSA is considered "no regulation" too if the authors start with the idea of exemption from liability as no regulation. Second, it is precisely regulation (Section 230/DSA) that protects the content platforms from *further* regulation. I agree with the idea behind the authors describing this as "no regulation", but the fact that they then introduce specific regulation contradicts this framing.

O2: Section 2 offers a clear overview of the legal avenues for content moderation and how the Digital Services Act fits within these avenues. In addition, the methodology of the Digital Services Act is clearly set out by the authors, in broadly understandable legal terms. I would ask the authors to consider, however, to remove explicit references to article numbers where this is not required for understanding the legal rules. While these article numbers might prove necessary for a strictly legal audience, they might unnecessarily distract a broader audience. I would also refrain as much as possible from using direct citations from the legal text; instead, the authors, who show excellent legal knowledge, could “translate” the legal rules to fit within the overall argument (e.g., the quote at the end of section 2.3 refers to “Union law or the law of any Member State”, which might be unclear to certain readers).

O3: In Section 2, the part where the authors compare the DSA notion of ‘accuracy’ with its counterpart in the AI Act requires revision. The authors also retrieve their inter-instrumental knowledge quite indirectly: starting with the AI Act, to High Level Expert Group on AI guidelines, to a complementary footnote therein. The authors write quite extensively about this comparison, which could essentially be reduced to a short reference to the guidelines of the High Level Expert Group on AI. I would ask the authors, therefore, to consider reducing this section.

O4: In the conclusion, the authors ask for administrative guidelines or legislative changes to improve the role of "accuracy" in the Digital Services Act. The DSA has already been enacted into law and is currently in force. If the paper was offered during the legislative process of the

DSA, the call for legislative changes would have been more applicable. The authors might want to reconsider how they provide this recommendation.
----------- Miscellaneous Minor Issues -----------
"Art. 15(1)(e) DSA only asks for "indicators of the accuracy and the possible rate of error" (Art. 15(1)(e) DSA)." Here, the authors mention the specific article number twice in the same sentence. This small problem is indicative of how the very explicit mentions of article numbers throughout Section 2 makes the text more difficult to read.

----------------------- REVIEW 2 ---------------------
SUBMISSION: 121
TITLE: Operationalizing content moderation ``accuracy'' in the Digital Services Act
AUTHORS: Johnny Wei, Frederike Zufall and Robin Jia

----------- Summary -----------
This paper deals with clarifying the term "accuracy" as introduced in the DSA when it comes to content moderation. Authors expand their analysis with the AI act to reflect on the use of term "accuracy", where more clarifications are provided as to what we mean by accuracy. Then, they embark on a journey on how performance of content moderation should be measured and provide a reflection on true positives, test data and reflect on metrics like accuracy, precision and recall. Authors provide a framework on how to use recall and stratified sampling to operationalize this on actual classifiers.
----------- Contributions -----------
(B) Critique or analysis of an existing problem framing or existing approach to a solution
----------- Relevance: The submission has relevance to FAccT -----------
SCORE: 2 (agree)
----------- Overall evaluation Score -----------
SCORE: -1 (weak reject: Marginally below acceptance threshold. I tend to vote for rejecting this submission, but accepting it would not be that bad.)
----------- Reviewer's confidence -----------
SCORE: 4 (high: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.)
----------- Review -----------
S1: Nice transition from current legal framework to the operationalization of DSA/moderation.
S2: Clear recommendations for platforms in the conclusion

O1: Focusing just on recall is really limiting the impact of the work. Why not reflect on other metrics as well? AUROC, which balances both? In sensitive domains (e.g. health) recall is of course vital, but this is not justified properly in the paper.
O2: The methodology is not properly justified: Why stratified sampling and binning etc. is the best approach here?

O3: The results lack interperation. Authors seem to target the # of samples to reach the prevalence of toxicity, but what about the performance of the model in terms of model "accuracy"? How does that come into play?
O4: While not the main point, Roberta models might be considered not state-of-the-art for the problem we are discussing here.

While the discussion of how a content moderation system can be implemented is very relevant (see summary), I believe that the paper ignores fundamental questions about how such systems will work in practice and how they are going to be evaluated, thus it cannot be accepted at its current form. The points O1-O4 can improve the paper.

----------------------- REVIEW 3 ---------------------
SUBMISSION: 121
TITLE: Operationalizing content moderation ``accuracy'' in the Digital Services Act
AUTHORS: Johnny Wei, Frederike Zufall and Robin Jia

----------- Summary -----------
The new DSA legislation in Europe imposes several obligations on online platforms related to content moderation. One of which is to provide ”indicators of accuracy and error rate” if they use automated algorithms to detect non-compliant content.

The paper argues that the DSA is not precise enough in what it means by accuracy, and discusses several metrics of accuracy. The paper also proposes to use stratified sampling to report accuracy and it shows its applicability on a Reddit dataset.
----------- Contributions -----------
(A) Identification or articulation of a novel problem, phenomenon, or potential societal harm;
----------- Relevance: The submission has relevance to FAccT -----------
SCORE: 3 (strongly agree)
----------- Overall evaluation Score -----------
SCORE: 0 (borderline paper: Marginally above acceptance threshold; an accept. I tend to vote for accepting this submission, but rejecting it would not be that bad.)
----------- Reviewer's confidence -----------
SCORE: 4 (high: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.)
----------- Review -----------
Understanding the DSA and the obligations it imposes is important for the community, so thank you for doing research in this space. I really enjoyed the description of the different obligations with respect to content moderation.

However, I think the central assumption the paper makes is wrong: the paper considers that DSA asks platforms for the “accuracy” (as a metric) of their methods and argues that “accuracy” might not be a good metric to report. From my understanding of the text the DSA asks for “indicators of accuracy” and not “accuracy” as a metric. Indicators of accuracy could well be precision, recall or other metrics.

The paper also presents different accuracy metrics that could be reported. For a computer scientists this is basic knowledge and the paper does not bring something new.

The paper also proposes to use stratified sampling when reporting accuracy. While this could be a useful tool in the toolbox for platforms, it is far from being the only one, and the paper only considers a simplified view of the content moderation

The paper does not seem to understand all the complications of content moderation and the fact that stratified sampling might be relevant for toxicity analysis, but not in other scenarios such as political speech detection where ground truth is more subjective.

Having said this, I do think the paper will generate useful discussions at the conference.

S1 New important problem on how the applicability of the DSA.
S2 Will generate discussions on the DSA.

O1 The paper should tone down the claims that the DSA is wrong in not having a mode precise description of what “indicators of accuracy” is. Given how complex is the content moderation problem, having precise definition would create more harm than good as one particular metric would not be meaningful across different scenarios.