Towards More Granular Reasoning: Mapping the DSA Transparency Database Taxonomy to Platform Terms and Conditions Using Large Language Models

Alyson RANUCCI [a,b], Eline LEIJTEN [b] and David GRAUS [a]

[a] *University of Amsterdam, The Netherlands*

[b] *Vrije Universiteit Amsterdam, The Netherlands*

ORCiD ID: Alyson Ranucci https://orcid.org/0009-0005-2077-8412, Eline Leijten https://orcid.org/0009-0006-9917-2372, David Graus https://orcid.org/0000-0002-6245-0870

Abstract. Article 24(5) of the Digital Services Act (DSA) requires providers of online platforms to submit every content moderation decision they make, with a corresponding Statement of Reasons (SoR), to the DSA Transparency Database (DSA-TDB). Such decisions are classified against seventeen high-level attribute categories that categorize the primary reason for the moderation decision. Because over 99.8% of statements cite a violation of the provider's own terms and conditions (which is a residual high-level category attribute choice) rather than a legal provision, the database is overwhelmingly a record of contractual enforcement using catch-all reporting; however, whether its categories actually describe those contractual rules well has not been examined. In this paper, we map the terms and conditions (T&C) of 31 designated online services, grouped into six service types, against the DSA-TDB's high-level category taxonomy using a few-shot GPT-4.1 classifier over 447 policy documents. The analysis concerns how T&C rules are represented in the taxonomy, not how they are enforced.

Two findings emerge. First, high-level category applicability varies sharply by service type: unsafe products account for 17.5% of e-commerce assignments and 0.2% of adult-platform assignments, and negative effects on civic discourse reaches 9.9% within the Meta family but is absent from adult platforms, Booking.com, and Wikipedia. Second, the categories have notable overlap among sub-category specifications, which makes assigning a single high-level category to a decision difficult. For instance, one of the residual catch-all categories accounts for 21.4% of assignments and is often assigned alongside a substantive category; the overlap is also visible in the published schema itself. We argue that catch-all reporting is plausibly in part a consequence of the DSA-TDB's schema and attribute design, and recommend a schema that is differentiated by service type with more mandatory attributes derived inductively from platform governance language.

Keywords. Digital Services Act, Transparency Database, Statement of Reasons, Terms and Conditions, Content Moderation

1. Introduction

Providers of online services make an enormous number of content moderation decisions every day, each marking an exercise of private power over a user's communication. Users frequently perceive enforcement as arbitrary and unevenly applied [1], and the resulting mistrust extends to platform governance as a whole [2].

The Digital Services Act (DSA) responds directly to user mistrust by offering greater transparency into online governance. Article 17 requires providers of hosting services to submit a Statement of Reasons (SoR) to any recipient whose content has been restricted, which must be *"clear and easily comprehensible and as precise and specific as reasonably possible under the given circumstances"* [3]. Article 24(5) requires providers of online platforms to submit every SoR to a publicly accessible database; as a result, the European Commission created the DSA Transparency Database (DSA-TDB), which, since February 2024, has accumulated a continuously updated record of moderation actions across all designated services. Each submission conforms to a fixed schema whose most consequential field is a mandatory high-level category attribute with seventeen values that include animal welfare, protection of minors, and illegal or harmful speech. Three of the categories are residual (i.e., catch-all groupings) and the remaining fourteen are substantive. Because the moderated content itself is excluded for privacy reasons, the category label is the primary signal available to users, regulators and researchers about what kind of rule was enforced when the moderation decision was made.

A substantial body of empirical work reports that the DSA-TDB delivers formal availability of moderation data without substantive transparency. Optional attributes intended to support granular reporting go largely unused [4]. In addition, free-text explanations are frequently formulaic and machine-generated [5], nearly half of analyzed platforms supply no meaningful information beyond the mandatory minimal categorical label [6], and the resulting reporting mechanism has not produced interoperable data that can be compared against other reporting mechanisms [7]. The prevailing assessment is one of formal compliance without substantive insight [8].

Current literature has examined the reports and decisions, but not the categories they are classified against. Kaushal et al. [5] estimate that over 99.8% of submitted statements cite a violation of the provider's own terms rather than any legal provision. The DSA-TDB can therefore overwhelmingly be considered a record of how platforms apply their own contractual rules, and whether its categories describe those rules well is a question about platforms' terms and conditions. Additionally, the same seventeen categories apply to every designated provider irrespective of what the service does, so a social network, a marketplace, a search engine, an encyclopedia and an adult content site all use an identical list to categorize their moderation decisions.

We therefore ask: *to what extent do the high-level categories used to classify Statements of Reasons correspond to the content governance rules that providers set out in their terms and conditions, and does that correspondence hold uniformly across service types?* We answer this question by mapping the terms and conditions of 31 designated Very Large Online Platforms (VLOPs) and Search Engines (VLOSEs) onto the DSA-TDB high-level attribute taxonomy, using a few-shot LLM classifier over policy documents. Our analysis concerns regulatory representation, i.e., how contractual rules are classified by the taxonomy, and not moderation behavior.

The paper makes three contributions. First, it provides the first systematic mapping between the DSA-TDB taxonomy and the contractual instruments of a large number of

designated providers; prior work linking statements to contractual text covers only individual platforms or statements. Second, it presents evidence that the taxonomy does not discriminate cleanly over contractual text, including a residual value that behaves as a competing substantive category. The core of this finding is visible in the published schema independently of any classifier or annotator, while the classifier-derived result is weaker and presented with the caveats its validation warrants. Third, it argues for differentiating the schema by service type and for the addition of more mandatory attributes to record the kind of rule enforced rather than only an alleged harm.

2. Background and Related Work

2.1. The Obligation, the Schema, and the Contractual Basis of Moderation

Art. 3(t) DSA defines *content moderation* as activities taken by online providers to address alleged illegality of content or incompatibility with the provider's own terms and conditions [3]. Art. 17 DSA requires that an SoR identify which ground applies and explain why it was relied upon, thereby presupposing a traceable correspondence between a moderation action and a clause of the provider's rules. The governing standard outlined in Art. 17(4) DSA, however, is purely evaluative; *as precise and specific as reasonably possible* has no benchmark or interpretive guidance, so the required granularity is delegated to the regulated providers.

Submissions conform to a schema of 35 attributes of which two are of notable importance: (i) the decision ground, and (ii) the high-level category [9]. The high-level category is a single value chosen from an exhaustive list of seventeen values, of which fourteen name a type of harmful or illegal content and three are residual. Two further attributes, `category_addition` and `category_specification`, are optional. The high-level category values vary in their specificity; some are narrowly delimited, such as *Negative effects on civic discourse* and *Animal welfare*, while others are broad, such as *Violence*. Of the three residual categories, one is a catch-all for any unspecified breach of the provider's terms; however, the use of this term does not require any additional clarification of the contractual rules that have been violated.

Despite such a lack of clarity, Kaushal et al. [5] report that over 99.8% of SoRs cite this category, meaning that contractual grounds act as the effective default. They also state that the contractual grounds are easier to invoke than statutory frameworks that may vary across Member States, which plausibly explains the asymmetry of contractual grounds used over legal grounds. Similarly, Aspromonte et al. [10] observe that the resulting communication of reasoning for moderation decisions done by online providers is typically generic, meaning that most of what providers submit to the database merely cites a violation of terms and conditions but nothing more.

2.2. Reporting Granularity and Its Limitations

Though a large number of online providers submit daily SoRs to the DSA-TDB, their use varies greatly, which further complicates how moderation decisions are explained. Groesch et al. [6] document inconsistent formats and incomplete reporting; Drolsbach and Pröllochs [11] report that TikTok performs nearly 350 times more moderation ac-

tions per user than X without corresponding differences in the depth of reasoning; and Trujillo et al. [4] find the optional granularity attributes seldom used. After the Implementing Regulation on transparency reporting [12], they find platforms still apply divergent procedures and recommend that further attributes be made mandatory [7]. Wei et al. [13] and Sekwenz et al. [14] attribute such patterns to the DSA's reliance on undefined evaluative concepts. Since the moderated content is excluded, and researcher access under Art. 40 DSA has not proved reliable [15], the remaining public resource is the provider's own terms and conditions, which motivates our approach.

2.3. Computational Analysis of Contractual Text

CLAUDETTE applies machine learning to detect potentially unfair clauses in terms of service against a legally motivated taxonomy [16], an approach since extended to multilingual settings [17]. This work is similar to ours in that it aligns heterogeneous contractual language with a fixed set of legal categories. Two studies address the SoR-T&C relationship directly: Aspromonte et al. [10] use large language models to explain SoRs by reference to retrieved sections of the applicable terms; Eßer and Spanakis [18] link TikTok's statements to the most relevant clause, annotate 100 statements, and conclude that both are frequently broad, many amounting to catch-all phrasing. Both map statements to their explaining text for a single platform; we invert and generalize this, mapping the taxonomy itself onto the terms of all designated providers. We also draw on work in legal ontologies, where the granularity of a schema and the relation between its general and specific classes are themselves objects of analysis [19,20]; the DSA-TDB taxonomy is a flat list of harms, applied to rules that are not all harm-based.

3. Method

3.1. Corpus

We analyze terms and conditions (T&Cs) of 31 online services designated as Very Large Online Platforms (VLOPs) or Very Large Online Search Engines (VLOSEs), from the European Commission's DSA Terms and Conditions Database (T&C Database) [21], a repository of versioned terms and conditions documents of these platforms. Where the DSA-TDB records the decisions providers report, the T&C Database records the documents that define the rules those decisions claim to enforce.

Retrieval from the T&C Database yielded 480 documents in July 2026; we filtered for English documents, since we validate the classification for English text, and filtered out purely procedural ones, leaving 447 documents. We group the 31 services into six types: (i) the Meta family (5 services), (ii) social media (6), (iii) e-commerce (6), (iv) search and general technology (8), (v) adult platforms (4), and (vi) other (Booking.com and Wikipedia). Document counts are imbalanced, e.g., TikTok Shop has 47 documents and Meta AI only one, so we do not compare aggregate rates between individual platforms.

3.2. Chunking and Classification

T&C documents are chunked hierarchically: the Markdown files are split at headings, so that each chunk belongs to a single policy section and carries a `section_path` as metadata. The body text is then split at paragraph boundaries and greedily accumulated into chunks up to a limit of 8,000 characters. The classifier (`gpt-4.1`) compares each chunk against all 17 categories (3 of which are residual and 14 of which are substantive categories) and returns (i) a match type, (ii) a confidence score, (iii) verbatim evidence quotes, and (iv) a reasoning summary. The prompt supplies the DSA-TDB categories with their official sub-category specifications, and a set of 20 few-shot examples from real T&C documents, with no training or parameter update. No cross-platform labeled resource that links T&C provisions to DSA-TDB categories exists (the closest covers 100 statements of a single platform [18]), which is why we apply few-shot learning. The match type only distinguishes between `explicit_match` and `no_match`; each match requires an evidence quote (a section from the T&C chunk that may have led the model to select a certain match type) and a reasoning summary to help the model ground its classification in the source document. Text that merely states that the platform follows a law, describes moderation abstractly, or explains how to report content is to be labeled `no_match` unless it names prohibited acts or restricted content. The classification yields 6,225 chunk-level high-level category assignments, from a single model and prompt configuration (Section 6). The prompt, few-shot examples, chunking code, model outputs and annotated sample are available from the authors on request.

3.3. Evaluation Design

We evaluate against 75 manually annotated chunks, stratified by the classifier's own prediction: 60 chunks with at least one category assigned, and 15 with none. At a corpus-level positive rate below 4% of chunk $\times$ category cells, random sampling would yield too few positives to estimate precision, while sampling only predicted positives would make recall unmeasurable. The annotator, one of the authors without formal legal training, recorded the categories each chunk substantively addresses, blind to the model's labels but aware of the study's framing. With a single annotator, inter-annotator agreement cannot be computed; we treat it as a single-reader sanity check, not a gold standard, and no central claim is stated as annotator-validated. Scoring is per chunk $\times$ category (1,050 binary predictions over 14 substantive categories); we report precision, recall and F1 rather than accuracy, since `no_match` labels would dominate in this heavily imbalanced scenario [13]. Since the sample is stratified on the classifier's output, figures are sample-level and not corpus-level estimates, and the 15-chunk negative stratum supports little inference about recall. We therefore use the classifier as an exploratory instrument for descriptive purposes and not as a reliable category-level measure.

4. Results

4.1. Where the Classifier and the Annotator Disagree

First, at the chunk level (whether a passage addresses any category), recall is 1 and precision 0.617. In the evaluation, the annotator and classifier agree that 37 chunks address

at least one category, and 15 none. However, in 23 cases the classifier assigns a category, whereas the annotator assigns none. The annotator finds in *no* chunks a category that the classifier misses. The above rests on only 15 predicted-negative chunks and is weak evidence that the classifier rarely misses text that is relevant to the high-level categories.

Second, at the assignment level, where we consider each chunk–label pair a separate prediction, precision for `explicit_match` is 0.298, recall 0.425 and F1 0.351 (accuracy, at 0.940, is uninformative as only 3.8% of cells are positive matches). Precision below recall means the classifier assigns categories more readily than the annotator, so reported coverage is an upper bound, and since the error directions are independent, absence is interpretable only where the classifier recognizes a category elsewhere in the corpus. Precision is highest for categories naming specific conduct, e.g., *Negative effects on civic discourse* 0.75, *Unsafe products* 0.67, and lowest for the residual *Other violation of provider's terms and conditions* (0.12), which absorbs 23 of the 65 false positives. The disagreement admits three readings our design cannot separate: taxonomy indeterminacy, classifier error, and an unreliable single annotation. We therefore weigh evidence from the published schema most heavily and treat classifier-derived patterns as suggestive.

4.2. The Categories Do Not Cleanly Partition Governance Text

Evidence from the published schema. The strongest evidence does not depend on any classifier or annotator. For instance, the residual category *Other violation of provider's terms and conditions* has sub-category specifications that include *Adult sexual material*, *Nudity*, *Age-specific restrictions* and *Goods/services not permitted to be offered on the platform*. These duplicate the subject matter of the high-level categories *Protection of minors* (which specifies *age-specific restrictions concerning minors*), *Unsafe, non-compliant or prohibited products*, and *Cyber violence*. On the face of the schema, the residual competes with the substantive categories for material they are defined to describe. Moreover, every substantive prohibition in a set of terms is also a violation of those terms, so a provider choosing between a substantive category and the residual chooses between two true descriptions with no rule of choice.

Evidence from the classifier's output. Figure 1 reports how often categories are assigned to the same chunk. A cleanly discriminating taxonomy would produce a strongly diagonal matrix. The observed matrix is not: *Scams* co-occurs with *Data protection* 216 times, and with *Unsafe products* 201; *Intellectual property* with *Data protection* 198; and *Protection of minors* with *Unsafe products* 183. These counts come from the classifier alone and, given its over-assignment, are likely inflated; they are an upper bound, consistent with but not independent confirmation of the overlap visible in the schema.

The most heavily co-assigned value is a residual category. *Other violation of provider's terms and conditions* is assigned 1,330 times, 21.4% of the total, and is the only value present on all 31 platforms. It co-occurs with *Scams* 352 times, *Intellectual property* 313 and *Unsafe products* 263. However, a residual meant for use when no substantive category is relevant should not behave this way. The other two residual values do behave as residuals; *Not captured by any high-level category* is assigned 31 times and co-occurs once with the catch-all, and *Type of alleged illegal content not specified* is assigned five times. Of the uncorroborated assignments to the residual, the annotator mostly assigned a different substantive category; however, these observations also rest on a single annotator since the classifier was not precise for the residual categories.

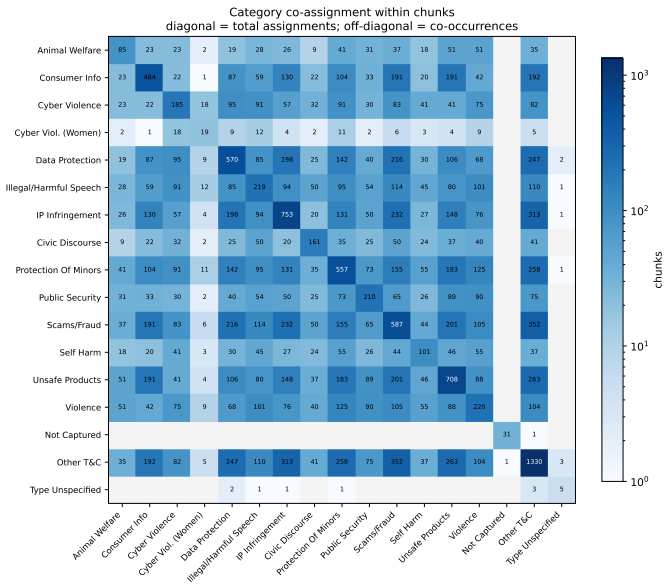

Figure 1. Category co-assignment within chunks. Diagonal cells = total assignments per category; off-diagonal cells = co-occurrences. Gray indicates no co-assignment.

Cyber violence against women is a limiting case. Of its 19 assignments, 18 also carry *Cyber violence*. The category is a subset of its parent, distinguished only by the gender specification, even though platform terms are drafted in gender-neutral language. For example, a Stripchat provision addressing exactly the conduct the category names drew the parent category because nothing in the text supports the narrower one by explicitly addressing women.

4.3. Coverage

Table 1 reports total assignments per category. No category goes unused, and the median platform is credited with coverage of 12 of the 14 substantive categories; given over-assignment, these are upper bounds. The problem is not that categories are dead but that they are unevenly and, we suggest, ambiguously filled, from 1,330 assignments for the residual to 19 for *Cyber violence against women*. Coverage correlates moderately with corpus size ($r = 0.43$), but Wikipedia, with nine documents (comparable to Instagram or Pinterest), yields text assigned to only two substantive categories; the categories absent from its terms (minors, violence, scams) are each assigned on more than 18 platforms elsewhere, so their absence is unlikely to be a detection failure. For an encyclopedia, this suggests that the categories it must report against largely do not correspond to its own policy instruments.

4.4. Variation by Service Type

Table 2 reports the share of each service type's substantive assignments in selected categories, and shows large differences. *Unsafe products* accounts for 17.5% of e-commerce assignments and 0.2% of adult-platform assignments; *Consumer information* ranges

Table 1. Chunk-level assignments per category (total 6,225). †residual value.

Category	Assign.	Category	Assign.
Other violation of T&Cs†	1,330	Violence	220
Intellectual property	753	Illegal or harmful speech	219
Unsafe or prohibited products	708	Public security	210
Scams and/or fraud	587	Cyber violence	185
Data protection and privacy	570	Civic discourse or elections	161
Protection of minors	557	Self-harm	101
Consumer information	484	Animal welfare	85
Not captured†	31	Cyber violence against women	19
Type not specified†	5		

Table 2. Share (%) of each service type's substantive assignments, with selected categories. Extremes in bold.

Category	Meta	Social	E-com.	Search	Adult	Other
Data protection	13.5	9.8	10.8	**15.0**	13.9	**22.5**
Unsafe products	8.8	**15.8**	**20.5**	7.2	**0.4**	0.8
Consumer information	4.0	6.3	15.2	8.1	**0.7**	**28.7**
Protection of minors	13.3	13.4	7.8	9.8	**25.6**	10.9
Civic discourse	**15.6**	4.8	0.4	2.5	**0.0**	**0.0**
Assignments (n)	377	1,625	1,806	641	281	129

from 22.0% for the "Other" group (Booking.com and Wikipedia) to 1.8% for adult platforms; and *Negative effects on civic discourse* reaches 9.9% in the Meta family, but is absent from adult platforms, Booking.com and Wikipedia, which reflects the content rather than a limit on what the classifier can recognize. *Data protection* is evenly distributed (16.8–24.5%), plausibly because it is backed by a mature compliance framework (the GDPR) that standardizes privacy terms across services. Groupings are small, so these are descriptive contrasts and not tested differences. Additionally, the levels may be inflated by over-assignment, though that does not by itself explain a category being assigned some 87 times more readily for one service type than another.

5. Discussion

5.1. A Taxonomy of Harms Applied to Documents That Are Not Only About Harm

Fourteen of the seventeen values name a type of harmful or illegal content, yet under Art. 3(u) DSA, terms and conditions encompass all clauses governing the contractual relationship, including seller obligations, advertising formats, account eligibility, monetization conditions and developer access. In the current taxonomy, there is no value for a decision grounded in a seller-verification requirement or an account-eligibility condition. This mismatch is a likely contributor to the pattern in Section 4.2: the residual is one of the few values for the body of platform rules that do not concern a harm, and

appears to be applied both to provisions its own specifications claim and to commercial or procedural rules for which the taxonomy offers no alternative.

5.2. Catch-All Reporting and Schema Design

The finding that 99.8% of submitted statements are dominated by catch-all labels is ordinarily read as platform behavior: providers taking the least demanding route to compliance. However, our results suggest this may not be the full story. A classifier with no compliance burden or litigation exposure, given the same seventeen options and documents, also classifies more text with the catch-all category than any substantive category. Over-assignment inflates the level and a model-specific bias cannot be ruled out, but neither obviously explains why the inflation lands on this single category, and the overlap in the published specifications offers an independent reason to expect it. Catch-all reporting is therefore plausibly in part a consequence of a design that asks providers to classify contractual provisions against a taxonomy of harms, to an undefined standard of specificity, with informative fields made optional [4,5,6,7]. However, this does not exclude provider incentives, since our design compares two textual layers and cannot test behavior. Nor do we treat the classifier-annotator disagreement as showing that human readers would diverge; it is equally consistent with model or annotation error.

5.3. Toward a More Granular Schema

We propose four revisions, motivated chiefly by the schema-level evidence and prior literature; they are hypotheses to test, not conclusions established by our classifier. First, a second, orthogonal *mandatory* attribute recording the *kind* of rule enforced, e.g., harm-based, commercial, procedural or eligibility-based, so that harm categories apply only where a harm is at issue, rather than forcing a choice between an ill-fitting harm value and a residual that conveys nothing. Its values should be derived inductively from platform governance language and validated with methods from legal ontology engineering [19, 20]; this study can pose that question but not settle it.

Second, service-type differentiation: service-specific values *added to* a preserved common core, attached to the rule-type attribute, so cross-type comparison is unaffected while within-type comparison becomes possible. The objection that this cuts against the harmonization sought by the Implementing Regulation [12] is met in part by Trujillo et al. [7], who find platforms still applying divergent procedures after harmonization, concluding that specifications were clear but compliance was not. The trade-off is between a measured cost, up to 21.4% of assignments in a value that conveys little and providers such as Wikipedia reporting against fourteen categories of which two appear to describe their rules, and a speculative one that a preserved core largely avoids.

Third, informative fields should move from optional to mandatory. For example, `category_addition` and `category_specification` already offer granularity but providers do not use them [4]; Trujillo et al. [7] reach the same recommendation from the opposite end of the pipeline. Fourth, and at lowest cost, the sub-category specifications require differentiation: faced with the overlap shown in Section 4.2, the broader catch-all may be the safer choice for a provider exposed to enforcement risk, so distinguishing the categories in their specifications, not merely their names, is a precondition for the rest.

6. Limitations

First, the study compares two textual layers and says nothing about enforcement, though a correspondence shows that a provider has articulated a rule addressing a category and not that it is enforced or consistently applied; Wagner et al. [22] found written terms accounted for only a fraction of one platform's removals. Our conclusions concern the representation of rules in the taxonomy and not moderation behavior.

Second, the reference annotation was produced once, by the first author without legal training, who was aware of the study's framing, so inter-annotator agreement could not be computed and the disagreement in Section 4.1 may reflect annotation or classifier error rather than taxonomy ambiguity. Two lines of evidence are not vulnerable to annotation error: the duplication among sub-category specifications is a property of the published schema, and the co-assignment structure derives from the classifier alone, albeit subject to its over-assignment; the paper's claims are weighted accordingly. Measuring how much of a corpus a category covers would require several legally trained annotators and reported agreement.

Third, the 75-chunk sample is stratified by the classifier's predictions, so figures are sample-level, recall rests on 15 negative chunks, and precision is low overall (0.298) and for the residual (0.12).

Fourth, all results derive from one proprietary model and prompt, untested against other models, prompts or annotation strategies.

Fifth, the analysis is restricted to English-language text and a July 2026 snapshot.

Finally, the mapping is a structured textual similarity judgment, not a legal determination; it is adequate for exploratory comparison but not for establishing what a provision legally covers.

7. Conclusion

We asked whether the categories used to classify Statements of Reasons correspond to the governance rules providers set out in their terms and conditions, and whether that holds across service types. Our evidence points toward a negative answer on both parts, with differing levels of strength. The overlap between the residual's specifications and substantive categories is observable in the published schema, whereas the classifier-derived patterns are suggestive and limited by modest agreement with a single annotator. The problem is not that categories go unused, since every category is assigned somewhere, but that the taxonomy and the documents it is applied to appear to concern partly different subject matter. Moreover, the taxonomy's applicability varies by service type even though the reporting obligation does not, and its values overlap enough that assigning a provision to one may often be a choice among plausible options rather than an explicit determination.

The DSA-TDB was intended to make moderation decisions legible to the users they affect. However, our results suggest that the known catch-all reporting may follow in part from poor schema design and not solely from unwillingness on the part of providers, though, since we analyze representation and not enforcement, we cannot say how much each factor explains. Better reporting is therefore unlikely to come from stricter enforcement of the present schema alone; it may more likely come from a schema that asks providers questions that their own terms can answer.

Acknowledgements

David Graus is partly funded by ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

References

[1] Duffy BE, Meisner C. Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility. Media, Culture & Society. 2023;45(2):285-304. Available from: https://doi.org/10.1177/01634437221111923.

[2] Savolainen L. The shadow banning controversy: perceived governance and algorithmic folklore. Media, Culture & Society. 2022;44(6):1091-109. Available from: https://doi.org/10.1177/01634437221077174.

[3] European Parliament and Council of the European Union. Regulation (EU) 2022/2065 of 19 October 2022 on a single market for digital services (Digital Services Act); 2022. Official Journal of the European Union, L 277, pp. 1–102.

[4] Trujillo A, Fagni T, Cresci S. The DSA Transparency Database: Auditing Self-reported Moderation Actions by Social Media. Proc ACM Hum-Comput Interact. 2025 May;9(2). Available from: https://doi.org/10.1145/3711085.

[5] Kaushal R, Van De Kerkhof J, Goanta C, Spanakis G, Iamnitchi A. Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. FAccT '24. New York, NY, USA: Association for Computing Machinery; 2024. p. 1121–1132. Available from: https://doi.org/10.1145/3630106.3658960.

[6] Groesch S, Birrer A, Just N, Saurwein F. Big data, small answers: How the DSA Transparency Database falls short of its regulatory objectives. Telecommunications Policy. 2026;50(1):103088. Available from: https://www.sciencedirect.com/science/article/pii/S0308596125001855.

[7] Trujillo A, Tessa B, Cresci S. Disarranged harmonization of transparency reporting by social media platforms under the Digital Services Act; 2026. Preprint.

[8] Klafkowska-Waśniowska K, Drąg J. In search of reason: Are statements of reason under the Digital Services Act a valuable source of information on content moderation practices?; 2025. SSRN Working Paper. Available from: https://ssrn.com/abstract=5257655.

[9] European Commission. DSA transparency database: API and schema documentation; 2024. Accessed: 2026-09-13. https://transparency.dsa.ec.europa.eu/page/api-documentation.

[10] Aspromonte M, Ferraris A, Galli F, Contissa G. LLMs to the Rescue: Explaining DSA Statements of Reason with Platform's Terms of Services. In: Aletras N, Chalkidis I, Barrett L, Goanță C, Preoțiuc-Pietro D, Spanakis G, editors. Proceedings of the Natural Legal Language Processing Workshop 2024. Miami, FL, USA: Association for Computational Linguistics; 2024. p. 205-15. Available from: https://aclanthology.org/2024.nllp-1.17/.

[11] Drolsbach CP, Pröllochs N. Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database. In: Companion Proceedings of the ACM Web Conference 2024. WWW '24. New York, NY, USA: Association for Computing Machinery; 2024. p. 939–942. Available from: https://doi.org/10.1145/3589335.3651482.

[12] European Commission. Commission Implementing Regulation (EU) 2024/2835 of November 2024 laying down templates concerning the transparency reporting obligations of providers of intermediary services and of providers of online platforms under Regulation (EU) 2022/2065; 2024. Official Journal of the European Union, L 2024/2835.

[13] Wei JTZ, Zufall F, Jia R. Operationalizing Content Moderation "Accuracy" in the Digital Services Act. In: Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society. AIES '24. AAAI Press; 2025. p. 1527–1538.

[14] Sekwenz MT, Wagner B, Parkin S. "It is unfair, and it would be unwise to expect the user to know the law!" – Evaluating reporting mechanisms under the Digital Services Act. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. FAccT '25. New York, NY, USA: Association for Computing Machinery; 2025. p. 532–546. Available from: https://doi.org/10.1145/3715275.3732036.

[15] Leerssen P. Outside the Black Box: From Algorithmic Transparency to Platform Observability in the Digital Services Act. Weizenbaum Journal of the Digital Society. 2024 May;4(2). Available from: https://ojs.weizenbaum-institut.de/index.php/wjds/article/view/4_2_3.

[16] Lippi M, Pałka P, Contissa G, Lagioia F, Micklitz HW, Sartor G, et al. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. Artificial Intelligence and Law. 2019 Jun;27(2):117-39. Available from: https://doi.org/10.1007/s10506-019-09243-2.

[17] Galassi A, Lagioia F, Jabłonowska A, Lippi M. Unfair clause detection in terms of service across multiple languages. Artificial Intelligence and Law. 2025 Sep;33(3):641-89. Available from: https://doi.org/10.1007/s10506-024-09398-7.

[18] Eßer L, Spanakis G. Linking Transparency and Accountability: Analysing The Connection Between TikTok's Terms of Service and Moderation Decisions. In: Aletras N, Chalkidis I, Barrett L, Goanță C, Preoțiuc-Pietro D, Spanakis G, editors. Proceedings of the Natural Legal Language Processing Workshop 2025. Suzhou, China: Association for Computational Linguistics; 2025. p. 382-95. Available from: https://aclanthology.org/2025.nllp-1.28/.

[19] Hoekstra R, Breuker J, Di Bello M, Boer A. The LKIF Core ontology of basic legal concepts. In: Casanovas P, Biasiotti MA, Francesconi E, Sagri MT, editors. Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2007). vol. 321 of CEUR Workshop Proceedings; 2007. p. 43-63.

[20] Sartor G, Casanovas P, Biasiotti M, Fernández-Barrera M, editors. Approaches to Legal Ontologies: Theories, Domains, Methodologies. vol. 1 of Law, Governance and Technology Series. 1st ed. Dordrecht: Springer; 2011.

[21] European Commission. Online platforms terms and conditions database; 2025. Accessed: July 2026. https://platform-contracts.digital-strategy.ec.europa.eu/.

[22] Wagner B, Kettemann MC, Tiedeke AS, Rachinger F, Sekwenz MT. Mapping interpretations of the law in online content moderation in Germany. Computer Law & Security Review. 2024;55:106054. Available from: https://www.sciencedirect.com/science/article/pii/S0267364924001201.