Opaque Transparency: Gaps and Discrepancies in the Report of Social Media Harms

Tyler Chang
College of Computing and Informatics
Drexel University
Philadelphia, Pennsylvania, USA
tlc373@drexel.edu

Joseph J Trybala III
Drexel University
Philadelphia, Pennsylvania, USA
jjt336@drexel.edu

Sharon Bassan
The Louis Brandeis Institute for Society, Economy, and Democracy
The College of Management Academic Studies
Rishon LeZion, Israel
sharonba4@colman.ac.il

Afsaneh Razi
Department of Information Science, College of Computing & Informatics
Drexel University
Philadelphia, Pennsylvania, USA
afsaneh.razi@drexel.edu

Abstract

Social media transparency reports exist as an eclectic collection of documents and data files that underdeliver on their advertised transparency and lack a shared lexicon of relevant harms, keeping many of the crucial details obfuscated from users. Although previous research has identified some of the harm categories that are underexplained by or absent from the reports, much of this work did not conform the enumerated subject areas into an easily digestible format. Through a comparative analysis of the reports and established sociotechnical and algorithmic harm taxonomies, we elucidate the gaps in the reporting of harm on social media and highlight the reports' inaccessibility to most users. We demonstrate a lack of discussion of particular harm categories, such as the environmental costs, data sales practices, legal obligations, and limitations on platforms' self-moderation, and propose a nutrition label for transparency that enables users to inform themselves about the relevant social media harms.

CCS Concepts

• **Human-centered computing** → **Social media**.

Keywords

social media, transparency report, harm taxonomy, nutrition labels

ACM Reference Format:
Tyler Chang, Joseph J Trybala III, Sharon Bassan, and Afsaneh Razi. 2025. Opaque Transparency: Gaps and Discrepancies in the Report of Social Media Harms. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3706599.3719829

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3719829

1 Introduction

With an estimated 4.9 billion users worldwide [41], social media is capable of exposing people to ideas, connections, and cultures otherwise inaccessible to them, while also exacerbating some forms of harm including sexual harassment and abuse [27, 46, 54, 55], security threats [26, 29, 64], mis- and disinformation [18, 58], and violent extremist ideologies [38, 51]. According to the National Center for Missing & Exploited Children (NCMEC), there have been over 195 million reports of digital child sexual abuse material (CSAM) since 1998 [7]. Furthermore, a 2024 survey found that 22% of Americans "experienced severe harassment on social media in the past 12 months" [12], a 4% increase over the previous year. Beyond these individuals-focused harms, large-scale data breaches have occurred semi-frequently [29, 30, 64].

A resource available to social media users, researchers, policymakers, and other direct and indirect stakeholders to learn more about these online risks and their frequencies is the transparency reports published by many social media companies. User-focused laws such as the EU's General Data Protection Regulation (GDPR) [1] and Data Services Act (DSA) [10, 34], and Germany's Network Enforcement Act (NetzDG) [2] mandate social media companies disclose a lengthy list of statistics about their platforms. However, significant problems persist with social media regulation. While these laws have meted out fines totaling billions of euros to social media companies [37], except for child safety topics (CSAM/CSEA), there is little that is consistently discussed in the reports [15].

Previous research has highlighted the insufficiency and sometimes absence of defined terminologies [15, 44, 57], verifiable statistics [6, 15], legal requirements [34, 56, 61], access to data for researchers (APIs) [15], and justifications for not providing a report [6]. Unfortunately for the interested user, these reports seldom deliver the desired insights in a readily interpretable form [15, 44, 56, 57, 62]. While previous research found these inconsistencies, they did not provide a comprehensive view of the categories of harm reported. Also, the content and formats of these reports vary considerably both between platforms and year-over-year. We therefore seek to answer the following questions: **RQ1:** *What are the trends, gaps, and discrepancies in the reporting of harms in the transparency reports of popular social media platforms?* **RQ2:** *How*

can we effectively communicate the harms being risked when using social media to users?

Therefore, we employed qualitative content analysis to compare the harm categories discussed in 12 popular social medias' transparency reports against established algorithmic and sociotechnical harm taxonomies. In answer to RQ1, we found that few, if any, reports discussed the increasing environmental costs associated with the platforms' use of generative AI, the inherent limitations of their abilities to moderate their own platforms, or user data selling practices. We formulated a plain language categorization of the relevant harm types and argue that a higher percentage of harm categories appearing in a report does not imply a better report. In answer to RQ2, we argue that the optimal solution is not merely further refinement of the reports but the introduction of a standardized, modular, and plain language-based nutrition label for transparency to enable users to quickly identify information about they harms they care about.

This work makes the following novel research contributions: (1) a comprehensive overview of the coverage of harm categories by the transparency reports, (2) a standard against which the quality of future reports can be evaluated, and (3) a novel visualization of the core information about transparency that is interpretable to expert and non-expert users.

2 Background

This section discusses researchers' examinations of the transparency reports, the three harm taxonomies against which we evaluated the reports, and the concept of a technical nutrition label.

2.1 Previous Research on Social Media Transparency

Researchers and organizations have evaluated transparency reports on how comprehensive, consistent, regular, easily found, and verifiable the reports were. A 2024 report from Anti-Defamation League(ADL) argued that the transparency reports provided only contextless metrics with little-to-no verifiability, little-to-no reporting of response times and other moderation details, utilized apparently deliberate unclear language, were often hard to find, and conveyed a decreased openness to providing researchers with access to APIs and other forms of data access [15]. A later report discussed these same platforms in refer to Californian law AB 587, finding that while all platforms either complied with the law's requirements or gave reasons as to why they did not qualify as a "social media", several did so under explicit protest [15]. Platforms like YouTube, which claims not to be a social media [6], and Mastodon, which asserts that effective moderation is impossible because of its decentralized structure [65], have capitalized on the absence of well-defined legal regulations in the US and EU to provide only obfuscated reports. Schaffner et al. expanded on these findings in their examination of 43 online platforms' handling of harmful speech, misleading content, and copyright infringement [57]. Both the ADL and Schaffner et al. found that vague terms such as "anti-social" often appeared in the reports without definitions and that the year-over-year changes to the reports and available resources made verification of much of the reports' content excessively difficult if not impossible [6, 15, 57]. There is a growing trend against free access to resources that support the evaluation of transparency, including X and Reddit ending free access to their data for researchers in 2023 and Meta replacing CrowdTangle with a less transparent alternative [15]. This is expected to accelerate the growth of misinformation on social media [18, 58].

In response to this poor state of affairs, researchers have proposed several modifications and additions to the transparency reports such as the recommendation that the platforms release their source code to the public and academic researchers [52]. Tool such as Microsoft PhotoDNA [13] and Google CSAI Match [9], which have already seen limited adoption by social media platforms, can be leveraged to better monitor for CSAM/CSEA. Michal Luria, has argued that recommendation systems often obscure their mechanisms to the detriment of users and suggested that the inclusion of some details of their mechanisms in the reports could make them more user-centric [49]. Atreja et al. have proposed AppealMod, a system that aims to reduce moderators' workloads and exposure to toxic material by mandating the inclusion of more context from users prior to a human moderator reviewing an appeal [28]. Relatedly, Song et al. designed ModSandbox, a virtual environment for testing moderation rules prior to their deployment with real users and detecting false positives and negatives in automated moderation systems [60]. Though these suggestions hold significant merit, they do neither address the environmental costs nor resolve the lack of communicative clarity for users. Moreover, while past research has investigated harm categorization, none have yet to compare the reports' categories against established taxonomies.

2.2 Sociotechnical and Algorithmic Harm Taxonomies

We used three taxonomies in our analysis of the reports. The first, created by Shelby et al. in 2023, focuses on sociotechnical instead of purely algorithmic or technical harms [59]. It consists of 5 primary harm categories—representational, allocative, quality of service, interpersonal, social system—and 20 subcategories, and highlights the often difficult to describe relationship between social and technical harms. The second taxonomy by Abercrombie et al. is a plain language, high-level taxonomy of concrete, non-technical harms [25]. Consisting of 9 primary categories and 68 subcategories, it is designed as an evolving framework that is adaptable to emerging technologies and changing digital environments. The third and final taxonomy by Raji et al. focuses on the Functionality Fallacy, the oftentimes ignored problem of assuming a system predominantly works as intended without verifying the truth of the matter [53]. Composed of only 4 primary categories and 10 subcategories, it is both the shortest and most abstract of the taxonomies. These are not nearly the only applicable harm taxonomies. There are those that focus on privacy [50], generative AI and altered content [39, 46], dis- and misinformation [58], social harms [31], and other broad categorizations of harm [26, 32, 33]. To better focus our research on making transparency accessible to users, we emphasized common interpretability and simpler language in our selection of taxonomies.

2.3 Technical Nutrition Labels

Even the most precise categorization of social media harms is useless to users unless said categories are also understandable. In a 2009 paper, Kelley et al. proposed the idea of a nutrition label for privacy [42]. Writing about the lack of clarity in previous designs for communicating privacy topics to users, the authors highlight the presence of unclear terminology and difficulties in navigating the multiple sources users were required to visit to gain access to all relevant information, challenges that apply to social media transparency as well. Their proposed nutrition label for privacy utilized color coding and other visual cues to simplify the dense and highly technical details of digital privacy for websites, ultimately yielding a single image that conveyed much of the critical information in a commonly interpretable format. In a 2010 study, Kelley et al. also demonstrated that users react positively to standardized formats, citing improved speed and accuracy in finding the desired information[43]. This was expanded to discuss the privacy and security of Internet of Things (IoT) devices in 2020 by Emami-Naeini et al., who employed interviews with both experts and non-experts to identify areas of interest and revise the label design [35]. Their designs, which included both simplified and full versions, again used a mix of text and visual cues to convey the core information in a condensed format. Emami-Naeini et al. offered some support for the likelihood of such labels being implemented, noting that Finland has already mandated the inclusion of them for IoT devices [35, 36]. We propose here a preliminary nutrition label for transparency that captures the core information without demanding users parse large volumes of metric-ridden and often highly technical text.

3 Methods

We collected transparency reports for the most popular social media platforms (n=12) in the US [17] including X, Tumblr, Snapchat, LinkedIn, Meta (Facebook, Instagram), Google (YouTube), TikTok, Twitch, Reddit, BlueSky, and Mastodon from mid-2023 to the end of 2024. First, for a more detailed understanding of what harm categories the reports present and how they are defined, two authors reviewed the reports inductively and in discussions with all authors, summarized them in Figure 1. Second, each report was independently compared against each of the taxonomies deductively by the same two authors to determine whether the report discussed each taxonomy subcategory to identify missing subcategories from a more holistic view. A primary category was marked as being covered by the report if all its subcategories were covered. While we examined supplemental information (help centers, community guidelines), we limited ourselves to the reports themselves during the comparison process. If a report did not offer a definition for a term and it could not be reasonably assumed that the category addressed a subcategory, it was marked as missing. For example, BlueSky uses the phrase "ToS violation" (terms of service violation) as a harm category but does not define what this means in the report [4]. All discrepancies in the coded comparisons were discussed between the two authors until all conflicts were resolved. Based on these comparisons, we performed qualitative content analysis to identify primary and subcategories of harm that were not addressed by the reports. These gaps were then compared against those already established by other researchers. The preliminary label for transparency was originally created by the first author and revised based on discussion with other authors and researchers at the first author's institution.

	TikTok	Meta	Pinterest	X (Twitter)	Snapchat	Twitch	BlueSky	LinkedIn	Reddit	Google	Tumblr	Mastodon
Explicit Material	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✘	✘
Violent or disturbing content	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✘	✘
Hate Speech or discrimination	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✘	✘
Misinformation and disinformation	✔	✘	✔	✘	✔	✘	✔	✔	✘	✔	✘	✘
Impersonation and identity	✔	✔	✘	✔	✘	✘	✘	✔	✔	✘	✘	✘
Spam and malicious content	✔	✔	✔	✘	✔	✔	✔	✔	✔	✘	✘	✘
Harassment and Bullying	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✘	✘
Threats and Intimidation	✔	✔	✔	✔	✔	✘	✔	✘	✘	✘	✘	✘
Privacy and Security	✔	✘	✔	✔	✘	✔	✘	✘	✔	✘	✘	✘
Fraud and Scams	✔	✔	✘	✔	✘	✔	✘	✔	✘	✘	✘	✘
Community Guideline Violations	✔	✔	✔	✔	✔	✔	✔	✘	✔	✘	✘	✘
Child Safety	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✘	✘
Mental and Physical Health	✔	✔	✔	✔	✔	✔	✘	✘	✘	✔	✘	✘
Terrorism	✔	✔	✔	✔	✔	✔	✔	✘	✘	✔	✘	✘
Percentage	**100.0%**	**85.7%**	**85.7%**	**85.7%**	**78.6%**	**78.6%**	**71.4%**	**64.3%**	**64.3%**	**57.1%**	**0.0%**	**0.0%**

Figure 1: The Encoded Harm Categories

4 Results

4.1 Trends and Discrepancies in Reporting of Harm Categories

4.1.1 Formatting and Consistency of the Reports. By collecting transparency reports across two years, we are able to identify trends over time. Most obvious is that not all the platforms published reports in both years. Notably, X did not release a report in 2022 or 2023, BlueSky's 2023 report consisted of a single paragraph, Tumblr has not released a report since 2020, and Mastodon has never released a report. For these reasons, Tumblr and Mastodon are excluded from most of the subsequent figures.

The first transparency report, released by Google in 2010, predates the first regulatory law, NetzDG, by eight years. While most platforms have maintained consistent releases since starting, there are exceptions in Tumblr and X. The timing of release also varied, with reports being published as frequently as quarterly (Meta) and a rarely as annually [15]. For a full transparency report timeline, see Figure 5 in Appendix.

While some of the categories in Figure 1 are self-explanatory, others may be less so. Explicit material refers to nonviolent sexual or nude content, including adult pornography but excluding nude art. Community guideline violations is our term for harmful content that does not fit other categories. Child safety refers to CSAM, CSEA, and other forms of child endangerment. Mental and physical health broadly encompasses self-destructive behaviors and comments, including self-harm and suicide. Finally, terrorism includes the spreading of extremist ideas ("radicalization") and recruitment to violent extremist groups.

The most significant discrepancy in the reports is due to the lack of definitions of the main harm terminologies. X, for example, uses two different sets of terms to describe the harm categories in reference to their user conduct policy and their enforcement policy, but offers no guide to map from one set of terms to the other [22]. While some platforms store all relevant content under a single website domain, others split their content across multiple sites. For example, Tumblr's content is split between tumblr.com [19] and their parent company Automattic's site [3]. Since only a minority of platforms produce comprehensive reports, most reports refer to external documents or websites for definitions of terms, community

guidelines, and other policies. These supplemental materials do not follow the same release schedules as the transparency reports and, as a result, make unclear whether a policy described in supplemental documents applied during the period covered by the report.

4.1.2 Harm Types Reported. All platforms addressed explicit material, violent or disturbing content, hate speech or discrimination, harassment and bullying, and child safety in their transparency reports. We separated violent or disturbing content from threats and intimidation and terrorism on account of some platforms restricting, for example, videos of executions but not restricting conversations that were supportive of known extremist groups. Although most of the platforms group spam, scams, impersonation, and fraud as a single category, we divided them into three groups. Despite this, none of the segregated categories received attention from all platforms. Google, notably, missed all three [23]. Despite TikTok and Meta covering the highest percentages of the harm categories, they nonetheless produced reports that are neither sufficiently comprehensive nor commonly intelligible. For both platforms, an abundance of vague terms without full definitions was observed, forcing readers to assume the terms' meanings. By contrast, Reddit's report, despite scoring second lowest and not addressing misinformation, mental health, or terrorism, was among the more readable and consistent reports. Misinformation was generally underdefined, with only Pinterest dividing it into separate categories for medical, civic, and climate misinformation [24]. Finally, despite the growing usage of generative AI and its impact on exacerbating misinformation [18], there is no mention of generative AI in the reports except for TikTok and X, with the former referring to it as "Edited Media and AI-generated Content" [8] and the latter as a subcategory of sexual content [22].

4.2 Coverage of Harm Categories based on Established Taxonomies

An analysis of the reports in comparison to the three harm taxonomies showed that interpersonal (T1), autonomy (T2), physical (T2), psychological (T2), and societal & cultural (T2) harms were the only categories to be at least partly discussed by all platforms. Besides interpersonal harms, which received the greatest attention with 7 of the 12 platforms providing full coverage of the category, no other category had more than 3 platforms offer a full discussion. Of the 216 taxonomy category-platform combinations, 17 (7.87%) were fully covered, 84 were partly covered (38.89%), and 115 were not covered at all (53.24%). None of the platforms addressed all parts of any taxonomy, with the taxonomies seeing average coverage scores of 45.27%, 33.25%, and 14.58%, respectively. No category of any taxonomy received complete coverage from every platform.

Similar coverage scores did not imply that the same categories were missing from the reports. Meta and Pinterest, despite both scoring overall in the low-mid 50s, had considerably different performances with respect to taxonomy 1. While both scored highly in presentational harm (83.33%, 100%) and interpersonal harms (100%, 100%), they varied substantially in their coverage of allocative (100%, 0%) and social systems (0%, 60%) harms. Since the number of subcategories within each primary category varied, partial coverage for presentational harms with its five subcategories, for example, was more influential on the coverage score than allocative harms and its two subcategories. A similar effect was observed with psychological, human rights & civil liberties, and societal & cultural harms having the greatest impact for the second taxonomy and engineering and post-deployment failures for the third (see appendix for all figures).

No platform discussed the communication failures (T3) or environmental harms (T2), with allocative harms (T1), quality of service harms (T1), impossible tasks (T3), and engineering failures (T3) receiving only marginally greater attention. Taken together, these gaps represent a failure to discuss any overstated or misrepresented capabilities for the platforms, pollution, excessive use of natural resources, opportunity and financial loss for users, discrimination in providing services or necessitating greater labor for equal work without justification, theoretical and practical limitations to the moderation capabilities of the platforms, missing safety features, and failures to implement stated policies. This also includes the absence of any acknowledgment of the recommendation systems, data buying and selling practices, data breaches, and legal violations by the platforms.

We consider this analysis to be complementary to the findings presented in Section 4.1. The inductive analysis of the reports themselves enables us to examine how the reports compare against one another. By combining these findings with the analysis of the taxonomies, we capture not only the harm types already present in some of the reports but also relevant harm types that are not yet reported on.

4.3 Communicating Transparency to Users

The statistics in the transparency reports are mostly presented as raw totals. Even when a metric is displayed as a percentage, it is always in reference to the number of reports made by users. Given the public availability of the estimated number of users, we can instead render these statistics relative to the active user base. Pinterest, for example, receives an estimated 537 million users each month [45]. Instead of reporting only the 61,360 incidents reported of explicit material as a simple total [24], it could instead be represented as the proportion of reports per user, which is approximately 1 report per 8752 users. For platforms like Pinterest that already track how many users view content that violates their rules, thepercentage of the violating content never seen by users could be reported. Using explicit material as an example, 80% of the explicit material content was seen by users before its removal [24]. Many of the reports employ a combination of text, statistics, and graphics in their transparency reports. The graphics, most frequently bar and pie charts, are presumably intended to aid in interpreting the reports. They are often unaccompanied by a legend or explanation of their content. Like the vague language that pervades the discussions of harm categories, the reports' imagery lacks crucial details for achieving genuine transparency.

5 Discussion

5.1 No Perfect Reports

None of the 12 reports have all of the requirements for effective transparency. While platforms like SnapChat and Pinterest produce superior reports to those of LinkedIn or Google [15], they are far from perfect. The better reports avoid the terminological

	TikTok	Meta	Pinterest	X	Snapchat	Twitch	BlueSky	LinkedIn	Reddit	Google	Tumblr	Mastodon
Taxonomy 1	72.67	56.67	52.00	37.33	62.67	48.00	23.00	32.33	40.67	27.33	0.0	0.0
Taxonomy 2	52.34	29.93	57.62	29.01	51.99	25.94	22.24	24.79	19.68	18.92	0.0	0.0
Taxonomy 3	0.00	8.33	54.17	0.00	0.00	16.67	25.00	8.33	33.33	0.00	0.0	0.0

Figure 2: Percentage of Taxonomy Categories covered by Reports

Category Name	Platforms	Percentage
T2: Societal & cultural	10	83.0
T1: Interpersonal Harms	10	83.0
T2: Autonomy	10	83.0
T2: Physical	10	83.0
T2: Psychological	10	83.0
T1: Presentational Harms	9	75.0
T1: Social System Harms	9	75.0
T2: Human rights & civil liberties	9	75.0
T3: Post-deployment failures	5	42.0
T2: Political & economic	5	42.0
T2: Reputational	4	33.0
T2: Business & financial	3	25.0
T1: Allocative Harms	2	17.0
T3: Impossible tasks	2	17.0
T3: Engineering failures	2	17.0
T1: Quality of Service Harms	1	8.0
T2: Environmental	0	0.0
T3: Communication failures	0	0.0

Figure 3: Coverage of Taxonomy Harm Categories

inconsistencies and centralize their content under single domain names, but still suffer from nontrivial gaps in their content. Our findings further evidence the findings of the ADL [5, 6, 15] and Shaffner et al. [57]. The reports remain plagued by inconsistent content and unclear language. We expand beyond their work, noting the minimal discussion of the environmental costs involved in operating these platforms, the usage of generative AI, the practical and theoretical limitations of the platforms to moderate themselves, and data selling breaches and violations.

The topics of environmental harm and generative AI being used are connected, as the training and usage of generative AI carries an elevated environmental cost [63]. No platform acknowledges their own use of generative AI tools or the energy and financial expenditure required to utilize said tools. Given the number of users, this represents a substantial environmental cost which users may wish to be informed of prior to making use of generative AI features on social media. All platforms report harm as incidental rather than consequential. It is presumed that while harm may occur on a platform, it suffices to resolve the harms on a case-by-case basis instead of questioning the fundamental functionality of the platform itself. This is at best morally suspect and violates a reasonable standard of effort for protecting users. That said, we concede that total transparency is not a reasonable ask. There is proprietary and confidential information that is compromising to user safety and business competition. There are also inherent limitations to what can be moderated on social media. As pointed out by Mastodon Instance Administrators [65] and others [47, 62], the expectation for a platform's moderation efforts varies with its purpose and the jurisdictions within it operates. Reddit, for example, relies on community moderators as its primary means of moderation and permits a greater deal of misinformation and disinformation [20]. Fortunately, they include an abbreviated version of the moderator code of conduct in their transparency report [11], enabling users to better understand the moderation process. This sort of clarity helps elevate Reddit above its peers. Unfortunately, Reddit is also guilty of obfuscating their data selling practices [16] similar to other cases of data breaches and legal violations [14, 30, 37]. Users generally cannot opt out of having their data sold [15]. There is also the unresolved question of for whose benefit the reports are published. While laws mandating reports now exist, they did not for the first eight years following the first report being published by Google in 2010 [56]. The content of many of the current reports extends beyond the requirements of the GDPR, DSA, and AB 587. This, alongside the decreased access for researchers to APIs and the data required to verify the statistics mentioned in the reports, suggests that it is not for the benefit of researchers but for their users. To effectively communicate the core information about social media harm to users, however, a novel approach must be sought.

5.2 The Nutrition Label for Transparency

Transparency reports could educate users on the harms they risk being exposed to when using social media and consequently allow users to make well-informed decisions about using a platform. The current reports squander this potential. We propose here a preliminary nutrition label for transparency that is similar to those previously designed for privacy [35, 42, 48]. We suggest the following be included in the label: (1) release date of the label and the period to which the label refers, (2) the number of monthly users, (3) the country in which the company is headquartered, (4) how often the transparency reports are released, (5) likelihood of being exposed to each harm category, (6) data selling and buying practices, (7) moderation and policy enforcement methods, (8) any legal violations and data breaches that occurred during the reporting period, and (9) the percentage of harm categories addressed by the report. These proposed elements are based on a combination of our qualitative coding of the reports themselves and our analysis of the reports' qualities in comparison to the established harm taxonomies. Items 1-4 provide information about when, under what jurisdiction, and how many people are affected by the social media's policies . This enables users to quickly learn what

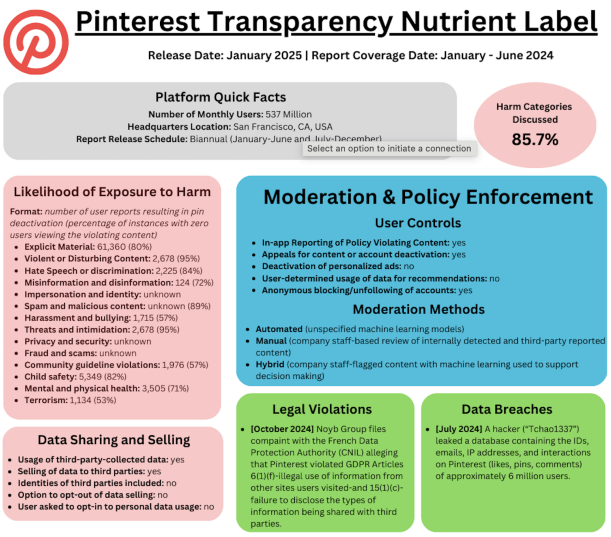

Figure 4: A Prototype of Transparency Label for Pinterest

rights they are afforded should they wish to report an instance of harm or appeal an automated decision made by the platform. For example, a Swiss user is entitled to a human review of any automated process under the nFADP [21] but the same is not assured for US residents, a contrast that the label would make clear. Item 5 addresses the issue of contextualizing the statistics found in the reports. As argued in Section 4.3, reframing these statistics will allow users to apply a more intuitive understanding of their relative risk of being harmed. Items 6 and 8 address the lack of user autonomy caused by not allowing users to opt-in or out of their data being sold and by omitting recent external dangers users may have been exposed to [25, 30, 40, 59, 64]. Item 7 aims to inform users about their options in the event of their experiencing harm or another form of dissatisfaction on social media. Moderation policies play a critical role in controlling what harms present greater or lesser risks for a given platform, and their being laid bare by the label will enable users to better evaluate the personal benefits and risks of their using any particular platform. Lastly, item 9, which is based on the qualitative coding of harms described in Figure 2, offers a single metric for users to make holistic comparatives between platforms. We envision the label using color-coding to group its components by general category (e.g., items 5 and 9 belong to the same group) and a modular design to allow for individual components to be added or removed in response to changes to the platforms. Though the details for each module will vary between platforms, the core structure of the label is intended to be platform-agnostic.

6 Limitations and future work

The nutrition label for transparency is an early stage effort and has not undergone the regiment of expert and non-expert feedback. Future work should conduct user studies such as surveys and interviews [35, 36, 48] to refine and validate the design of the transparency label. For many of the taxonomies that were not selected, the minutia of their designs was either too narrow to apply to a wide array of social media platforms or was too technical to be readily understood by users. That said, we concede that there is likely much of significance to be gleaned from evaluating specific dimensions of the reports against more narrowly defined taxonomies; it is simply beyond the scope of this article to do so. Similarly, our analysis included only 12 platforms and had the limited temporal scope of a single year. Expanding the analyses to platforms that are based outside of the United States or EU and considering a greater number of years' reports will enable future researchers to better identify how the reporting of harm on social media changes over time. By utilizing the three harms taxonomies for our analysis basis, we were able to identify the harms categories missing from the literature more objectively. Future researchers could employ different user-centered approaches to identify the priority of each types of harms to be presented in the transparency reports and the transparency labels proposed. While the implementation of these labels admittedly faces presently unresolved challenges, we nonetheless present them here as evidence for future research and of the merit of seeking their adoption by social media companies.

References

[1] 2018. The History of the General Data Protection Regulation | European Data Protection Supervisor. https://www.edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en
[2] 2021. Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz - NetzDG). https://perma.cc/RW47-95SR Section: Digital.
[3] 2021. Tumblr. https://transparency.automattic.com/tumblr/
[4] 2023. BlueSky 2023 Moderation Report. https://bsky.social/about/blog/01-16-2024-moderation-2023 Section: Digital.
[5] 2024. AB587 Revisited: How are Platforms Complying with California's Newly-Mandated Transparency Reporting? https://www.adl.org/resources/report/ab587-revisited-how-are-platforms-complying-californias-newly-mandated Section: Digital.
[6] 2024. ADL Finds Many Social Media Platforms Are Not Fully Complying with Newly Mandated Transparency Reporting. https://www.adl.org/resources/press-release/adl-finds-many-social-media-platforms-are-not-fully-complying-newly Section: Digital.
[7] 2024. Child Sexual Abuse Material. https://www.missingkids.org/theissues/csam Section: Digital.
[8] 2024. Community Guidelines Enforcement Report. https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2024-9 Section: Digital.
[9] 2024. Discovery our child safety toolkit. https://protectingchildren.google/tools-for-partners/ Section: Digital.
[10] 2024. The enforcement framework under the Digital Services Act. https://digital-strategy.ec.europa.eu/en/policies/dsa-enforcement Section: Digital.
[11] 2024. Moderator Code of Conduct. https://redditinc.com/policies/moderator-code-of-conduct Section: Digital.
[12] 2024. Online hate and harassment: The American experience. https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2024 Section: Digital.
[13] 2024. PhotoDNA. https://www.microsoft.com/en-us/photodna?oneroute=true Section: Digital.
[14] 2024. Pinterest Hack Affects Millions of Users. https://ai-techreport.com/massive-pinterest-hack-sounds-alarm-bells-for-millions-of-users
[15] 2024. Platform Transparency Reports - Just How Transparent? https://www.adl.org/resources/article/platform-transparency-reports-just-how-transparent Section: Digital.
[16] 2024. Reddit sells training data to unnamed AI company ahead of IPO. https://arstechnica.com/information-technology/2024/02/your-reddit-posts-may-train-ai-models-following-new-60-million-agreement/ Section: Digital.
[17] 2024. Social Media Fact Sheet. https://www.pewresearch.org/internet/fact-sheet/social-media/
[18] 2024. There's less social media transparency and, likely, more disinformation. https://thehill.com/opinion/technology/4881927-disinformation-social-media-transparency/ Section: Digital.
[19] 2024. Transparency Report. https://www.tumblr.com/transparency Section: Digital.

[20] 2024. Transparency Report: January to June 2024. https://redditinc.com/policies/transparency-report-january-to-june-2024 Section: Digital.

[21] 2024. Understanding the New Swiss Federal Act on Data Protection (FADP). https://secureprivacy.ai/blog/switzerland-new-federal-act-data-protection-fadp-key-changes-compliance Section: Digital.

[22] 2024. X Global Transparency Report. https://transparency.x.com/content/dam/transparency-twitter/2024/x-global-transparency-report-h1.pdf Section: Digital.

[23] 2024. YouTube Community Guidelines enforcement. https://transparencyreport.google.com/youtube-policy/removals?hl=en_GB Section: Digital.

[24] 2025. Transparency report. https://policy.pinterest.com/en-gb/transparency-report Section: Digital.

[25] Gavin Abercrombie, Djalel Benbouzid, Paolo Giudici, Delaram Golpayegani, Julio Hernandez, Pierre Noro, Harshvardhan Pandit, Eva Paraschou, Charlie Pownall, Jyoti Prajapati, Mark A. Sayre, Ushnish Sengupta, Arthit Suriyawongkul, Ruby Thelot, Sofia Vei, and Laura Waltersdorfer. 2024. A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms. https://doi.org/10.48550/arXiv.2407.01294 arXiv:2407.01294 [cs].

[26] Ioannis Agrafiotis, Jason R C Nurse, Michael Goldsmith, Sadie Creese, and David Upton. 2018. A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. *Journal of Cybersecurity* 4, 1 (Jan. 2018), tyy006. https://doi.org/10.1093/cybsec/tyy006

[27] Ashwaq Alsoubai, Afsaneh Razi, Zainab Agha, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. 2024. Profiling the Offline and Online Risk Experiences of Youth to Develop Targeted Interventions for Online Safety. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 114 (April 2024), 37 pages. https://doi.org/10.1145/3637391

[28] Shubham Atreja, Jane Im, Paul Resnick, and Libby Hemphill. 2024. AppealMod: Inducing Friction to Reduce Moderator Workload of Handling User Appeals. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 1–35. https://doi.org/10.1145/3637296

[29] Balaji. 2024. 6 Million Records of Pinterest Database Leaked - Cyber Press. https://cyberpress.org/6-million-records-of-pinterest-database-leaked/

[30] Guru Baran. 2024. Pinterest Data Leak: Hackers Claiming Access to 60 Million Rows of Data. https://cybersecuritynews.com/pinterest-data-leak/

[31] Andrew Critch and Stuart Russell. 2023. TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI. https://doi.org/10.48550/arXiv.2306.06924 arXiv:2306.06924 [cs].

[32] Douglas Cumming, Kumar Saurabh, Neelam Rani, and Parijat Upadhyay. 2024. Towards AI ethics-led sustainability frameworks and toolkits: Review and research agenda. *Journal of Sustainable Finance and Accounting* 1 (March 2024), 100003. https://doi.org/10.1016/j.josfa.2024.100003

[33] Alicia DeVrio, Motahhare Eslami, and Kenneth Holstein. 2024. Building, Shifting, & Employing Power: A Taxonomy of Responses From Below to Algorithmic Harm. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1093–1106. https://doi.org/10.1145/3630106.3658958

[34] Chiara Patricia Drolsbach and Nicolas Pröllochs. 2024. 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)*. Association for Computing Machinery, New York, NY, USA, 939–942. https://doi.org/10.1145/3589335.3651482

[35] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. 2020. Ask the Experts: What Should Be on an IoT Privacy and Security Label?. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Francisco, CA, USA, 447–464. https://doi.org/10.1109/SP40000.2020.00043

[36] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. 2022. An Informative Security and Privacy "Nutrition" Label for Internet of Things Devices. *IEEE Security & Privacy* 20, 2 (March 2022), 31–39. https://doi.org/10.1109/MSEC.2021.3132398

[37] CIPT Masha Komnenic CIPP/E FIP, CIPM. 2024. 61 Biggest GDPR Fines & Penalties So Far [2024 Update]. https://termly.io/resources/articles/biggest-gdpr-fines/

[38] Jarod Govers, Philip Feldman, Aaron Dant, and Panos Patros. 2023. Down the Rabbit Hole: Detecting Online Extremism, Radicalisation, and Politicised Hate Speech. *ACM Comput. Surv.* 55, 14s, Article 319 (July 2023), 35 pages. https://doi.org/10.1145/3583067

[39] Wiebke Hutiri, Orestis Papakyriakopoulos, and Alice Xiang. 2024. Not My Voice! A Taxonomy of Ethical and Safety Harms of Speech Generators. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 359–376. https://doi.org/10.1145/3630106.3658911

[40] Scott Ikeda. 2024. Pinterest Privacy Complaint Targets Platform for Familiar User Tracking Issues. https://www.cpomagazine.com/data-protection/pinterest-privacy-complaint-targets-platform-for-familiar-user-tracking-issues/

[41] Belle Wong J.D. 2023. Top Social Media Statistics And Trends Of 2025. https://www.forbes.com/advisor/business/social-media-statistics/ Section: Business.

[42] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. 2009. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security (SOUPS '09)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/1572532.1572538

[43] Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. 2010. Standardizing privacy notices: an online study of the nutrition label approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1573–1582. https://doi.org/10.1145/1753326.1753561

[44] Jordan Kraemer. 2024. On Social Media, Transparency Reporting is Anything But Transparent | TechPolicy.Press. https://techpolicy.press/on-social-media-transparency-reporting-is-anything-but-transparent

[45] Naveen Kumar. 2024. 28 Pinterest Statistics (2025) — Active Users Data. https://www.demandsage.com/pinterest-statistics/

[46] Hao-Ping (Hank) Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. 2024. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3613904.3642116

[47] Paddy Leerssen. 2020. The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems. *European Journal of Law and Technology* 11, 2 (Oct. 2020). https://ejlt.org/index.php/ejlt/article/view/786

[48] Tianshi Li, Kayla Reiman, Yuvraj Agarwal, Lorrie Faith Cranor, and Jason I. Hong. 2022. Understanding Challenges for Developers to Create Accurate Privacy Nutrition Labels. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–24. https://doi.org/10.1145/3491102.3502012

[49] Michal Luria. 2023. Co-Design Perspectives on Algorithm Transparency Reporting: Guidelines and Prototypes. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1076–1087. https://doi.org/10.1145/3593013.3594064

[50] Aleecia M Mcdonald and Lorrie Faith Cranor. [n. d.]. The Cost of Reading Privacy Policies. ([n. d.]).

[51] Muhammad Musa, Muhammad Usama, and Momin Uppal. 2023. Extremism on Social Media: Lynching of Priyantha Kumara Diyawadana. In *Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (Istanbul, Turkey) *(ASONAM '22)*. IEEE Press, 508–509. https://doi.org/10.1109/ASONAM55673.2022.10068622

[52] pklein. 2022. Transparency: The First Step to Fixing Social Media. https://ide.mit.edu/insights/transparency-the-first-step-to-fixing-social-media/

[53] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 959–972. https://doi.org/10.1145/3531146.3533158

[54] Afsaneh Razi, Ashwaq Alsoubai, Seunghyun Kim, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. 2023. Sliding into My DMs: Detecting Uncomfortable or Unsafe Sexual Risk Experiences within Instagram Direct Messages Grounded in the Perspective of Youth. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1 (April 2023), 89:1–89:29. https://doi.org/10.1145/3579522

[55] Afsaneh Razi, Karla Badillo-Urquiola, and Pamela J. Wisniewski. 2020. Let's Talk about Sext: How Adolescents Seek Support and Advice about Their Online Sexual Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376400

[56] Amanda Reid, Shanetta M. Pendleton, and Lightning (Joe) JM Czabovsky. 2024. Social media transparency reports: Longitudinal content analysis of news coverage. *The Journal of Social Media in Society* 13, 1 (May 2024), 122–154. https://thejsms.org/index.php/JSMS/article/view/1447 Number: 1.

[57] Brennan Schaffner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao Tan. 2024. "Community Guidelines Make this the Best Party on the Internet": An In-Depth Study of Online Platforms' Content Moderation Policies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3613904.3642333

[58] Connie Moon Sehat, Ryan Li, Peipei Nie, Tarunima Prabhakar, and Amy X. Zhang. 2024. Misinformation as a Harm: Structured Approaches for Fact-Checking Prioritization. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 171:1–171:36. https://doi.org/10.1145/3641010

[59] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 723–741. https://doi.org/10.1145/3600211.3604673

[60] Jean Y. Song, Sangwook Lee, Jisoo Lee, Mina Kim, and Juho Kim. 2023. ModSandbox: Facilitating Online Community Moderation Through Error Prediction and Improvement of Automated Rules. In *Proceedings of the 2023 CHI Conference*

on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3544548.3581057

[61] Caitlyn Vergara, Raghav Jain, and Swapneel Mehta. 2024. A History of Transparency Regulations: Interdisciplinary Strategies for Shaping Social Media Regulation and Self-Governance. In *Proceedings of the 25th Annual International Conference on Digital Government Research (dg.o '24)*. Association for Computing Machinery, New York, NY, USA, 875–883. https://doi.org/10.1145/3657054.3657157

[62] Ben Wagner, Krisztina Rozgonyi, Marie-Therese Sekwenz, Jennifer Cobbe, and Jatinder Singh. 2020. Regulating transparency?: Facebook, Twitter and the German Network Enforcement Act. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 261–271. https://doi.org/10.1145/3351095.3372856

[63] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. (2021). https://doi.org/10.48550/ARXIV.2112.04359 Publisher: [object Object] Version Number: 1.

[64] Davey Winder. [n. d.]. Warning As 26 Billion Records Leak: Dropbox, LinkedIn, Twitter Named. https://www.forbes.com/sites/daveywinder/2024/01/23/massive-26-billion-record-leak-dropbox-linkedin-twitterx-all-named/ Section: Cybersecurity.

[65] Zhilin Zhang, Jun Zhao, Ge Wang, Samantha-Kaye Johnston, George Chalhoub, Tala Ross, Diyi Liu, Claudine Tinsman, Rui Zhao, Max Van Kleek, and Nigel Shadbolt. 2024. Trouble in Paradise? Understanding Mastodon Admin's Motivations, Experiences, and Challenges Running Decentralised Social Media. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2 (Nov. 2024), 520:1–520:24. https://doi.org/10.1145/3687059

A Additional Figures

For Figures 9-11, a red X means no subcategories covered, an orange circle means partial coverage of the subcategories, and a green check mark means all subcategories covered.

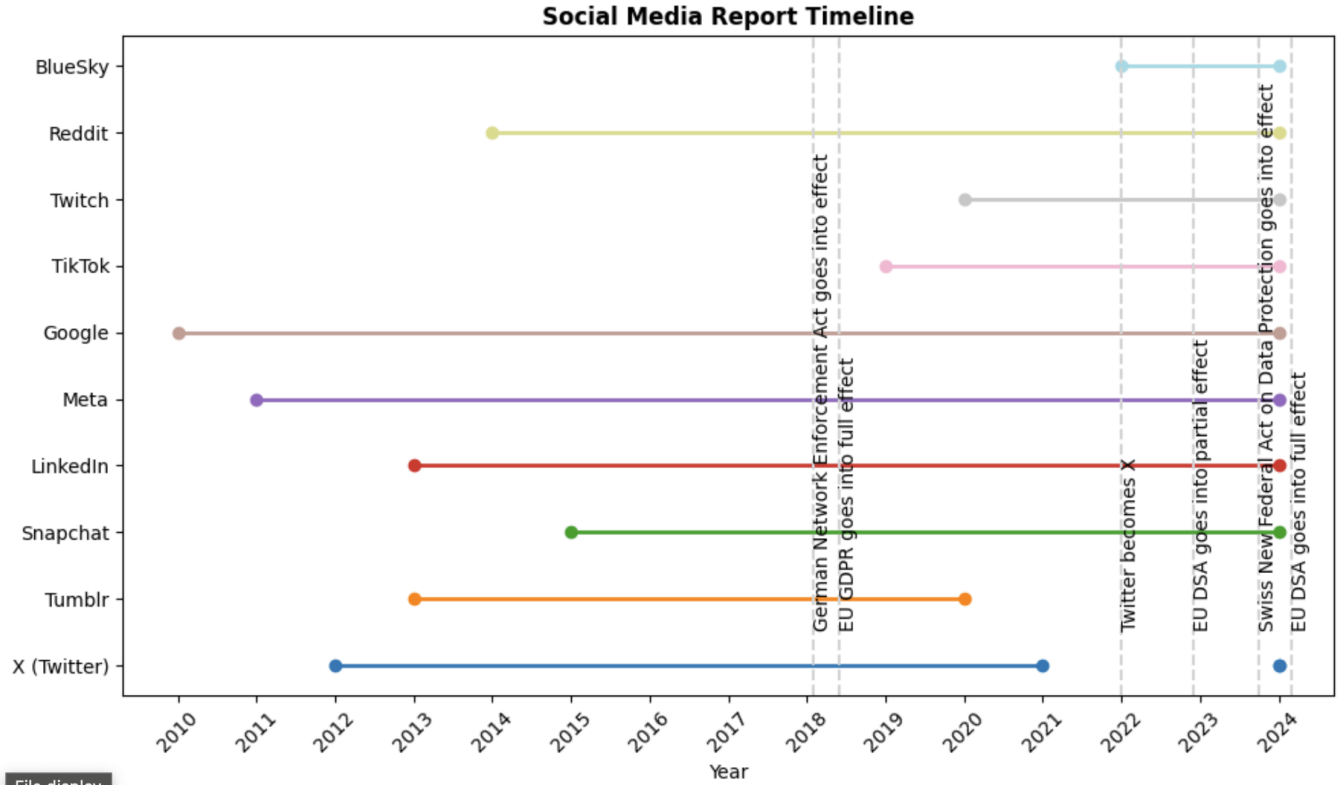

Figure 5: A Timeline of Transparency Reports being Produced

Taxonomy Source	Category Name	Primary Category	TikTok	Meta (Facebook/Instagram)	Pinterest	X (Twitter)	Snapchat	Twitch	BlueSky	LinkedIn	Reddit	Google (YouTube)	Tumblr	Mastodon
1	**Presentational Harms**	**none**	**83.33%**	**83.33%**	**100.00%**	**66.67%**	**100.00%**	**100.00%**	**0.00%**	**66.67%**	**83.33%**	**66.67%**	**0.00%**	**0.00%**
1	Stereotyping social groups	Representational harms	1	1	1	1	1	1	0	1	1	1	0	0
1	Demeaning social groups	Representational harms	1	1	1	1	1	1	0	1	1	1	0	0
1	Erasing social groups	Representational harms	1	1	1	1	1	1	0	1	1	1	0	0
1	Alienating social groups	Representational harms	1	1	1	1	1	1	0	1	1	1	0	0
1	Denying people the opportunity to self-identify	Representational harms	1	1	1	0	1	1	0	0	1	0	0	0
1	Reifying the essentialist social categories	Representational harms	0	0	1	0	1	1	0	0	0	0	0	0
1	**Allocative Harms**	**none**	**100.00%**	**100.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**
1	Opportunity loss	Allocative harms	1	1	0	0	0	0	0	0	0	0	0	0
1	Economic loss	Allocative harms	1	1	0	0	0	0	0	0	0	0	0	0
1	**Quality of Service Harms**	**none**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**33.33%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**
1	Alienation	Quality of service harms	0	0	0	0	1	0	0	0	0	0	0	0
1	Increased labor	Quality of service harms	0	0	0	0	0	0	0	0	0	0	0	0
1	Service/benefit loss	Quality of service harms	0	0	0	0	0	0	0	0	0	0	0	0
1	**Interpersonal Harms**	**none**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**75.00%**	**75.00%**	**100.00%**	**50.00%**	**0.00%**	**0.00%**
1	Loss of agency	Interpersonal harms	1	1	1	1	1	1	0	1	1	0	0	0
1	Tech-facilitated violence	Interpersonal harms	1	1	1	1	1	1	1	1	1	1	0	0
1	Diminished health and well-being	Interpersonal harms	1	1	1	1	1	1	1	1	1	1	0	0
1	Privacy violations	Interpersonal harms	1	1	1	1	1	1	1	0	1	0	0	0
1	**Social System Harms**	**none**	**80.00%**	**0.00%**	**60.00%**	**20.00%**	**80.00%**	**40.00%**	**40.00%**	**20.00%**	**20.00%**	**20.00%**	**0.00%**	**0.00%**
1	Information harms	Social system harms	1	0	1	1	1	1	1	1	1	1	0	0
1	Cultural harms	Social system harms	1	0	1	0	1	1	1	0	0	0	0	0
1	Civic and political harms	Social system harms	1	0	1	0	1	0	0	0	0	0	0	0
1	Socio-economic harms	Social system harms	1	0	0	0	1	0	0	0	0	0	0	0
1	Environmental harms	Social system harms	0	0	0	0	0	0	0	0	0	0	0	0
1	**Taxonomy 1**	**none**	**75.00%**	**55.00%**	**65.00%**	**45.00%**	**75.00%**	**60.00%**	**25.00%**	**40.00%**	**50.00%**	**35.00%**	**0.00%**	**0.00%**

Figure 6: Full comparison chart for the first taxonomy in comparison to the platforms

Taxonomy Source	Category Name	Primary Category	TikTok	Meta (Facebook/Instagram)	Pinterest	X (Twitter)	Snapchat	Twitch	BlueSky	LinkedIn	Reddit	Google (YouTube)	Tumblr	Mastodon
2	**Autonomy**	**none**	**75.00%**	**75.00%**	**100.00%**	**50.00%**	**100.00%**	**50.00%**	**25.00%**	**75.00%**	**75.00%**	**25.00%**	**0.00%**	**0.00%**
2	Autonomy/agency loss	Autonomy	0	0	1	0	1	1	0	1	1	0	0	0
2	impersonation/identity theft	Autonomy	1	1	1	1	1	1	1	1	1	1	0	0
2	IP/copyright loss	Autonomy	1	1	1	1	1	0	0	1	0	0	0	0
2	Personality rights loss	Autonomy	1	1	1	0	1	0	0	0	1	0	0	0
2	**Physical**	**none**	**75.00%**	**75.00%**	**75.00%**	**75.00%**	**75.00%**	**75.00%**	**50.00%**	**75.00%**	**50.00%**	**75.00%**	**0.00%**	**0.00%**
2	Bodily injury	Physical	1	1	1	1	1	1	1	1	1	1	0	0
2	Loss of life	Physical	1	1	1	1	1	1	1	1	1	1	0	0
2	Personal health detioration	Physical	1	1	1	1	1	1	0	1	0	1	0	0
2	Property damage	Physical	0	0	0	0	0	0	0	0	0	0	0	0
2	**Psychological**	**none**	**45.45%**	**54.55%**	**81.82%**	**54.55%**	**81.82%**	**54.55%**	**63.64%**	**27.27%**	**45.45%**	**54.55%**	**0.00%**	**0.00%**
2	Addiction	Psychological	1	0	1	0	0	0	0	0	0	0	0	0
2	Alienation/isolation	Psychological	0	0	0	0	1	0	1	0	0	0	0	0
2	Anxiety/distress	Psychological	0	0	1	0	1	0	0	0	0	0	0	0
2	Coercion/manipulation	Psychological	0	1	1	1	1	1	1	0	1	1	0	0
2	Dehumanisation/objectification	Psychological	0	1	1	1	1	1	1	0	0	1	0	0
2	Harassment/abuse/intimidation	Psychological	1	1	1	1	1	1	1	1	1	1	0	0
2	Over-reliance	Psychological	0	0	0	0	0	0	0	0	0	0	0	0
2	Radicalisation	Psychological	1	1	1	1	1	1	0	1	1	1	0	0
2	Self-harm	Psychological	1	1	1	1	1	1	1	1	1	1	0	0
2	Sexualisation	Psychological	1	1	1	1	1	1	1	1	1	1	0	0
2	Trauma	Psychological	0	0	1	0	1	0	0	0	0	0	0	0
2	**Reputational**	**none**	**50.00%**	**0.00%**	**50.00%**	**50.00%**	**50.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**
2	Defamation/libel/slander	Reputational	1	0	1	1	1	0	0	0	0	0	0	0
2	Loss of confidence/trust	Reputational	0	0	0	0	0	0	0	0	0	0	0	0
2	**Business & financial**	**none**	**33.33%**	**33.33%**	**16.67%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**
2	Business operations/infrastructure damage	Business & financial	0	0	0	0	0	0	0	0	0	0	0	0
2	Confidentiality loss	Business & financial	1	1	0	0	0	0	0	0	0	0	0	0
2	Financial/learnings loss	Business & financial	1	1	0	0	0	0	0	0	0	0	0	0
2	Livelihood loss	Business & financial	0	0	0	0	0	0	0	0	0	0	0	0
2	Monopolisation	Business & financial	0	0	0	0	0	0	0	0	0	0	0	0
2	Opportunity loss	Business & financial	0	0	1	0	0	0	0	0	0	0	0	0
2	**Human rights & civil liberties**	**none**	**81.82%**	**18.18%**	**63.64%**	**18.18%**	**36.36%**	**27.27%**	**27.27%**	**18.18%**	**0.00%**	**9.09%**	**0.00%**	**0.00%**
2	Benefits/entitlements loss	Human rights & civil liberties	0	0	0	0	0	0	0	0	0	0	0	0
2	Digity loss	Human rights & civil liberties	1	0	0	0	1	0	0	0	0	0	0	0
2	Discrimination	Human rights & civil liberties	1	1	1	1	1	1	1	0	0	0	0	0
2	Loss of freedom of speech/expression	Human rights & civil liberties	1	0	1	0	0	0	0	0	0	0	0	0
2	Loss of freedom of assembly/association	Human rights & civil liberties	1	0	0	0	0	0	0	0	0	0	0	0
2	Loss of social rights and access to public services	Human rights & civil liberties	1	0	0	0	0	0	0	0	0	0	0	0
2	Loss of right to information	Human rights & civil liberties	1	0	1	0	1	1	1	1	0	1	0	0
2	Loss of right to free elections	Human rights & civil liberties	1	0	1	0	0	0	0	0	0	0	0	0
2	Loss of right to liberty and security	Human rights & civil liberties	1	0	1	0	0	0	0	0	0	0	0	0
2	Loss of right to due process	Human rights & civil liberties	0	0	1	0	0	0	0	0	0	0	0	0
2	Privacy loss	Human rights & civil liberties	1	1	1	1	1	1	1	1	0	0	0	0
2	**Societal & cultural**	**none**	**53.33%**	**13.33%**	**60.00%**	**13.33%**	**53.33%**	**26.67%**	**20.00%**	**13.33%**	**6.67%**	**6.67%**	**0.00%**	**0.00%**
2	Breach of ethics/values/norms	Societal & cultural	1	0	0	0	0	0	1	0	0	0	0	0
2	Cheating/plagarism	Societal & cultural	1	0	0	0	0	0	0	0	0	0	0	0
2	Chilling effect	Societal & cultural	0	0	0	0	0	0	0	0	0	0	0	0
2	Cultural dispossession	Societal & cultural	0	0	1	0	1	1	0	0	0	0	0	0
2	Damage of public health	Societal & cultural	1	0	1	0	1	0	0	0	0	0	0	0
2	Historic revisionism	Societal & cultural	1	0	1	0	1	0	0	0	0	0	0	0
2	Information degradation	Societal & cultural	1	0	1	0	1	1	0	1	1	1	0	0
2	Job loss/losses	Societal & cultural	0	0	0	0	0	0	0	0	0	0	0	0
2	Labour exploitation	Societal & cultural	0	0	0	0	0	0	0	0	0	0	0	0
2	Loss of creativity/critical thinking	Societal & cultural	0	0	1	0	0	0	0	0	0	0	0	0
2	Stereotyping	Societal & cultural	1	1	1	1	1	1	0	1	0	0	0	0
2	Public service delivery detrioration	Societal & cultural	0	0	0	0	0	0	0	0	0	0	0	0
2	Societal destabilisation	Societal & cultural	1	0	1	0	1	0	1	0	0	0	0	0
2	Societal inequality	Societal & cultural	0	0	1	0	1	0	0	0	0	0	0	0
2	Violence/armed conflict	Societal & cultural	1	1	1	1	1	1	1	0	0	0	0	0
2	**Political & economic**	**none**	**57.14%**	**0.00%**	**71.43%**	**0.00%**	**71.43%**	**0.00%**	**14.29%**	**14.29%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**
2	Critical infrastructure damage	Political & economic	0	0	0	0	0	0	0	0	0	0	0	0
2	Economic instability	Political & economic	0	0	1	0	1	0	0	0	0	0	0	0
2	Power concentration	Political & economic	0	0	0	0	1	0	0	0	0	0	0	0
2	Electoral interference	Political & economic	1	0	1	0	1	0	0	0	0	0	0	0
2	Institutional trust loss	Political & economic	1	0	1	0	1	0	1	1	0	0	0	0
2	Political instability	Political & economic	1	0	1	0	0	0	0	0	0	0	0	0
2	Political manipulation	Political & economic	1	0	1	0	1	0	0	0	0	0	0	0
2	**Environmental**	**none**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**
2	Biodiversity loss	Environmental	0	0	0	0	0	0	0	0	0	0	0	0
2	Carbon emissions	Environmental	0	0	0	0	0	0	0	0	0	0	0	0
2	Electronic waste	Environmental	0	0	0	0	0	0	0	0	0	0	0	0
2	Excessive energy consumption	Environmental	0	0	0	0	0	0	0	0	0	0	0	0
2	Excessive landfill	Environmental	0	0	0	0	0	0	0	0	0	0	0	0
2	Excessive water consumption	Environmental	0	0	0	0	0	0	0	0	0	0	0	0
2	Natural resources extraction	Environmental	0	0	0	0	0	0	0	0	0	0	0	0
2	Pollution	Environmental	0	0	0	0	0	0	0	0	0	0	0	0
2	**Taxonomy 2**	**none**	**51.47%**	**26.47%**	**57.35%**	**23.53%**	**50.00%**	**26.47%**	**25.00%**	**20.59%**	**16.18%**	**17.65%**	**0.00%**	**0.00%**

Figure 7: Full comparison chart for the second taxonomy in comparison to the platforms

Taxonomy Source	Category Name	Primary Category	TikTok	Meta (Facebook/Instagram)	Pinterest	X (Twitter)	Snapchat	Twitch	BlueSky	LinkedIn	Reddit	Google (YouTube)	Tumblr	Mastodon
3	**Impossible tasks**	**none**	**0.00%**	**0.00%**	**50.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**100.00%**	**0.00%**	**0.00%**	**0.00%**
3	Conceptually impossible	Impossible tasks	0	0	0	0	0	0	0	0	1	0	0	0
3	Practically impossible	Impossible tasks	0	0	1	0	0	0	0	0	1	0	0	0
3	**Engineering failures**	**none**	**0.00%**	**0.00%**	**100.00%**	**0.00%**	**0.00%**	**0.00%**	**100.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**
3	Design failures	Engineering failures	0	0	1	0	0	0	1	0	0	0	0	0
3	Implementation failures	Engineering failures	0	0	1	0	0	0	1	0	0	0	0	0
3	Missing safety features	Engineering failures	0	0	1	0	0	0	1	0	0	0	0	0
3	**Post-deployment failures**	**none**	**0.00%**	**33.33%**	**66.67%**	**0.00%**	**0.00%**	**66.67%**	**0.00%**	**33.33%**	**33.33%**	**0.00%**	**0.00%**	**0.00%**
3	Robustness issues	Post-deployment failures	0	0	0	0	0	1	0	1	0	0	0	0
3	Failure under adversarial attacks	Post-deployment failures	0	1	1	0	0	0	0	0	0	0	0	0
3	Unanticipated interactions	Post-deployment failures	0	0	1	0	0	1	0	0	1	0	0	0
3	**Communication failures**	**none**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**
3	Falsified or overstated capabilities	Communication failures	0	0	0	0	0	0	0	0	0	0	0	0
3	Misrepresented capabilities	Communication failures	0	0	0	0	0	0	0	0	0	0	0	0
3	**Taxonomy 3**	**none**	**0.00%**	**10.00%**	**60.00%**	**0.00%**	**0.00%**	**20.00%**	**30.00%**	**10.00%**	**30.00%**	**0.00%**	**0.00%**	**0.00%**

Figure 8: Full comparison chart for the third taxonomy in comparison to the platforms

Category Name	TikTok	Meta	Pinterest	X	Snapchat	Twitch	BlueSky	LinkedIn	Reddit	Google	Tumblr	Mastodon
Presentational Harms	🟠	🟠	✅	🟠	✅	✅	❌	🟠	🟠	🟠	❌	❌
Allocative Harms	✅	✅	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌
Quality of Service Harms	❌	❌	❌	❌	🟠	❌	❌	❌	❌	❌	❌	❌
Interpersonal Harms	✅	✅	✅	✅	✅	✅	🟠	🟠	✅	🟠	❌	❌
Social System Harms	🟠	❌	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	❌	❌
Overall Performance	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	❌	❌

Figure 9: First Taxonomy in Comparison to the 12 Social Media Platforms

Category Name	TikTok	Meta	Pinterest	X	Snapchat	Twitch	BlueSky	LinkedIn	Reddit	Google	Tumblr	Mastodon
Autonomy	🟠	🟠	✅	🟠	✅	🟠	🟠	🟠	🟠	🟠	❌	❌
Physical	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	❌	❌
Psychological	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	❌	❌
Reputational	🟠	❌	🟠	🟠	🟠	❌	❌	❌	❌	❌	❌	❌
Business & financial	🟠	🟠	🟠	❌	❌	❌	❌	❌	❌	❌	❌	❌
Human rights & civil liberties	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	❌	🟠	❌	❌
Societal & cultural	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	❌	❌
Political & economic	🟠	❌	🟠	❌	🟠	❌	🟠	🟠	❌	❌	❌	❌
Environmental	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌
Overall Performance	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	❌	❌

Figure 10: Second Taxonomy in Comparison to the 12 Social Media Platforms

Category Name	TikTok	Meta	Pinterest	X	Snapchat	Twitch	BlueSky	LinkedIn	Reddit	Google	Tumblr	Mastodon
Impossible tasks	❌	❌	🟠	❌	❌	❌	❌	❌	✅	❌	❌	❌
Engineering failures	❌	❌	✅	❌	❌	❌	✅	❌	❌	❌	❌	❌
Post-deployment failures	❌	🟠	🟠	❌	❌	🟠	❌	🟠	🟠	❌	❌	❌
Communication failures	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌
Overall Performance	❌	🟠	🟠	❌	❌	🟠	🟠	🟠	🟠	❌	❌	❌

Figure 11: Third Taxonomy in Comparison to the 12 Social Media Platforms