SWBL KM - Research Paper

Digital Humanism or Empty Promises? A Longitudinal Analysis of Ethical Reporting in Model Cards

Zerda Polat

Student ID: 12323111

Subject Area: Information Business

Studienkennzahl: J123456789

Supervisor: Name of supervisor

Date of Submission: 30.June 2023

Department of Information Systems & Operations Management,
Vienna University of Economics and Business,
Welthandelsplatz 1, 1020 Vienna, Austria

Abstract

The governance framework for artificial intelligence (AI) systems frequently lags behind rigorous standards for Digital Humanism, which hinders algorithmic transparency and potentially enables unethical or biased deployment of systems. Currently, Model Cards (an instrument for standardized ethical reporting) have attracted considerable attention. Yet scarce evidence exists regarding the semantic quality of these cards. In this study, we empirically analyze the model documentation in the field and examine the content of ethical considerations within a curated dataset of 158 models. We find that the coverage of ethical risks is selective. While Major Tech actors prioritize social bias, they systematically omit ecological impacts. Thus, our study reveals the model of partial transparency facing current documentation practices and recommendations.

Keywords: Model Cards, Machine Learning Documentation, Digital Humanism, Algorithmic Transparency, AI Ethics, Greenwashing

1 Introduction

Ethical documentation serves as the primary resource to understand and evaluate opaque algorithmic systems when adopting them in critical domains. Machine-learned (ML) models increasingly are scrutinized as components of social infrastructure and would benefit from rigorous transparency (Mitchell et al., 2019; Bhat et al., 2023). Stakeholders resort to the documentation to answer questions such as what social biases are present, what limitations to expect, and what environmental and safety impacts to consider once the model is deployed at scale. Nevertheless, ML models shared as open-weights repositories or proprietary services are often inconsistently documented. Serious issues related to the opacity of ML models have been observed in various applications, notably in discriminatory hiring, biometric surveillance, and excessive energy consumption, leading to broader concerns about their alignment with Digital Humanism and social justice (Werthner et al., 2022).

As a reaction to observed problems in ML model transparency, significant efforts towards documenting models and datasets have been proposed. The Model Cards framework proposed by Mitchell et al. (2019) has become a standard, and the popular model hosting site HuggingFace has adopted the term model card. Yet, it is largely unknown how the semantic quality of these proposals has impacted the practice of documenting ML models and their undocumented costs, such as environmental impact.

In this work, we systematically study ML model documentation in the field and investigate how to encourage more responsible and holistic model documentation practice. While past work has shown limited documentation during model development (Bhat et al., 2023), we focus on the semantic content of ethical reporting for reusable ML models. We start by investigating how the ML models are documented, how they meet the ethical requirements within the model cards template proposed by Mitchell et al. (2019). Our study reveals that despite adopting the model card terminology, most model development teams fail to provide balanced and transparent documentation that can support scrutiny for ecological accountability. Certain aspects of documentation are especially limited across different contexts of model creators, such as information regarding the environmental impact.

We explore how we could assess the evolution of documentation standards and encourage truthful reporting practices. To this end, we analyze a curated dataset of 158 models, released between 2018 and 2025. A comparative analysis of these epochs demonstrates that, when Generative AI was introduced, data scientists adopted documentation approaches that prioritized

safety and misuse concerns over process transparency. They also showed less deliberate consideration of environmental costs compared to social biases. Our work makes the following contributions to understanding and supporting ML model documentation practice: (1) Delineation of the current practice of public model cards highlighting a clear gap between the social transparency provided by corporate actors and the ecological opacity maintained by the same developers; (2) A typology of ethical considerations, identified and used in our study based on the model cards proposal, which can be adopted by model auditors as a framework for detecting "greenwashing" or selective ethics; (3) A longitudinal analysis, covering the "Pre-Generative" and "Generative AI" eras, to support researchers to understand and critique the regression in documentation completeness during the recent industry expansion.

The artifacts created in this study including the coded dataset, list of assessed model cards, and analysis scripts, are shared as supplementary materials alongside the paper to support future investigation on improving ML documentation.

2 Related Work

2.1 Definitions and peculiarities of the problem

A difficult problem with evaluation in many real-world domains is that the performance of interest may depend on some hidden context, not given explicitly in the form of global accuracy. A typical example is facial recognition error rates that may vary radically with the demographic subgroup. Another example is the patterns of toxicity detection that may change with context, depending on the dialect, cultural background, or specific terminology used. Often the cause of failure is hidden, not known a priori, making the auditing task more complicated. Changes in the hidden context can induce more or less radical changes in the target reliability, which is generally known as bias (Mitchell et al., 2019). An effective documentation framework should be able to track such changes and to quickly expose them.

To validate this approach, Mitchell et al. (2019) applied the Model Cards framework to benchmarks such as the CelebA dataset, demonstrating that while models achieved high global accuracy, disaggregated reporting revealed error rates as high as 30% in specific subgroups. Thus, an ideal documentation system should be able to: (1) quickly adapt to subgroup performance; (2) be robust to aggregate metrics; and (3) recognize and treat hidden failure

modes. To this end, Mitchell et al. propose nine distinct sections to ensure accountability:

- **Model Details:** Lists basic metadata about the model, including the release date, version, architecture type, license, responsible parties, and citation information.
- **Intended Use:** Delineates the primary use cases and target users, as well as scenarios that are out-of-scope but easily confused with or highly related to the primary task.
- **Factors:** Records how demographic or phenotypic groups, alongside instrumental and environmental factors, influence model performance.
- **Metrics:** Covers the measurement of performance, including the specific thresholds, confidence intervals, and benchmarks utilized.
- **Evaluation Data:** Describes the datasets used to quantitatively evaluate the model, including justifications for dataset selection and any preprocessing procedures.
- **Training Data:** Details the provenance of the dataset used for training. When proprietary information cannot be disclosed, it should provide basic distributional statistics over groups.
- **Quantitative Analyses:** Illustrates performance through disaggregated evaluation with respect to the identified factors and their intersections.
- **Ethical Considerations:** Discusses the ethical considerations taken during development, such as the use of sensitive data, foreseen risks (e.g., environmental cost, bias), and mitigation strategies.
- **Caveats and Recommendations:** Lists additional concerns that are not covered in previous sections.

2.2 Types of documentation standards

Although previous conceptual frameworks and guidance for documentation have included the user perspective, key contextual and goal-related distinctions have not been fully discussed. To close this gap, the literature normally distinguishes between two kinds of documentation standards: (1) output-focused (Model Cards), and (2) lifecycle-focused (FactSheets).

The framework by Arnold et al. (2019), FactSheets, relies on theories of supplier conformity and highlights specific elements of the data lineage that

support auditing processes. Unlike Model Cards, which focus heavily on the final model output, FactSheets argue for documenting the entire lifecycle, including data provenance, cleaning, and testing methodologies. Although useful for identifying documentation elements to support specific safety-critical audits, this model shares a common limitation with other academic proposals: it relies on voluntary adherence by developers. This may lead to a lack of usability and practical interpretability of these documents for real end-users.

2.3 Systems for handling documentation practice

Prior work has tried to create categories of documentation to define transparency needs, but, as discussed in Bhat et al. (2023), practice often deviates from theory. Bhat et al. distinguish two approaches in available repositories: (1) technical metadata recording and (2) qualitative ethical discussion.

Current evaluation studies provide insight into how developers prioritize these approaches. Bhat et al. (2023) show in their experiments that qualitative ethical discussion handles transparency worse than analogous technical metadata techniques, which is likely due to the "Traceability Gap." Their empirical analysis showed that while technical metadata is almost always present, qualitative sections are frequently treated as noise. Specifically, they found that the "Ethical Considerations" section was missing or empty in a statistically significant portion of the sample. This implies that unless developers are explicitly nudged, they prioritize technical reproducibility over ethical transparency.

2.4 Base risks for handling environmental accountability

Researchers in the HCI and ML communities have proposed frameworks for user-centered design in documentation. This literature focuses mainly on who the documentation is provided to, the stakeholder, and why they require it,the specific verification goals. While these are important elements in understanding the context of use of a model, little attention seems to be paid to where or when users require documentation regarding ecological impact.

These questions relate to the environment in which a model is expected to operate. In the real world, risks may often be ecological, e.g., only particular types of carbon footprints may change with scale. In the case of Generative AI, many traditional templates are discarded simply because their coverage on current risks falls. The landscape of documentation requirements has been fundamentally altered by the advent of Foundation Models (2023–2025).

Bommasani et al. (2021) note that the move towards large-scale generative models has rendered the supply chain more opaque. Furthermore, Weidinger et al. (2021) argue that Large Language Models (LLMs) introduce specific risks, specifically highlighting environmental impact. They note that traditional documentation often fails to capture carbon footprint benchmarks. Motivated by these evolving risks, we seek to address whether current documentation practices have adapted to capture the ecological costs inherent to Generative AI.

3 Methodology Framework

3.1 Data Overview

The Hugging Face Hub is an open online repository that provides metadata for machine learning models, based on actual developers' uploads. It offers a variety of documentation sections, such as Model Details, Intended Use, and Ethical Considerations. Table 2 describes the features offered by the Model Cards framework and their respective descriptions.

When using the Hugging Face Hub for research, data are retrieved from the "Model Card" feature, which allows access to real-time metadata from 2018 up to the present. The data are retrieved directly from repository files in .json or markdown format after the examined model repositories are identified and the task, period, and creator type are selected. By default, the selection is filtered for models with significant download counts and citation history.

The data are stratified over the selected time frame as follows: Each data point is categorized by the model modality and the epoch it represents to compare relative documentation quality. Otherwise, domains with the most upload volume would always be ranked highest. The resulting observations are scaled based on a section's proportion to all documented ethical considerations.

The stratification indicates that results vary by institutional and temporal context. The value 0 for environmental reporting indicates very low disclosure volumes that are not included in the results. The process excludes placeholder entries that contain special characters or generic phrases. The Hub does not have an enforcement filter for ethical content, but it prompts users to complete sections related to bias and safety. It allows retrieval of documentation's normalized adherence for any model entered, independent of creator type.

The framework allows for various combinations to compare different terms and eras:

- For one theme in one creator type over a specific period, such as "Bias" in Major Tech Actors.
- For the same theme in different modalities, such as "Environmental Impact" in NLP and Computer Vision.
- For different themes in the same creator group, such as "Bias," "Safety," "Privacy," and "Environment" in Independent Developers.

When themes, modalities, periods, and creators are defined, the outputs are a distribution of the variations in documentation adherence over the selected time frame, presented separately for all examined domains; all datasets can be analyzed using descriptive statistics.

3.2 Keyword and Theme Selection

The selection of keywords when examining documentation text is key for valid results. The analysis is not case sensitive, but it takes into account plural forms and specific technical terminology. Therefore, parts of the respective documentation will not be considered if they utilize non-standard phrasing.

To partly overcome this, the iterative dictionary feature includes commonly encountered synonyms, selected and entered manually. For example, "environmental impact" is often described as "carbon footprint" or "energy consumption." Therefore, multiple terms can be entered as the search term by using dictionary expansion. In this way, results including standard and non-standard terms are aggregated.

In the case of "False Positives," variations in intent between terms without meaningful content should be explored. For example, "N/A" and "Not explicitly discussed" are searched for as negative phrases. In most cases, these are used as placeholders, and the analysis must exclude them.

Finally, when researching with Model Cards, the options of "technical metadata" and "ethical considerations" are available. It is imperative that theme selection is conducted with caution and that available dictionary features are analyzed to ensure validity.

3.3 Period and Era Selection

The selection of the examined time frame is a common requirement in longitudinal research. The main guideline is that the period selected for documentation data should be segmented based on the industry transition to Generative AI.

A single dataset was compiled including the years from 2018 to 2025. Depending on the time frame, the interval for which data are available varies significantly. The default selection includes the Pre-Generative Era and the Generative AI Era.

3.4 Threats to Validity

Finally, it is imperative that the scope of this framework is defined. Our content analysis identifies the presence of information rather than the technical correctness of reported metrics. Furthermore, we excluded linked external papers to focus on primary documentation within the repository. While the dataset of 158 models is robust, results are limited to English-language documentation. Nuanced discussions using non-standard terminology might be overlooked despite iterative dictionary expansion.

4 Results

The developers' difficulty in maintaining balanced documentation was a common feature of many repositories and, whilst the study had been designed expecting that this might be the case, the results often highlighted significant selective reporting, with both Major Tech and Independent actors being mistaken in their prioritization of ethical risks. We shall contrast the results between all 158 models, particularly with regard to establishing the difference between the corporate-backed disclosures and the independent disclosures, and comparing this to the difference between the Pre-Generative and Generative AI eras.

Of 158 models identified, 137 met inclusion criteria (86.71%), while the remaining either lacked ethical considerations sections or were left undocumented. In Table 1, we can clearly observe the extent of coverage and adoption organized by creator type.

In Fig. 7, we can clearly observe the distribution of model documentation organized by model creator and domain. In the diagram we can clearly observe two things:

Table 1: Extent of Coverage and Adoption by Creator Type (RQ1)

Category	**Count / %**	**Context**
Total Models	158	-
Documented	137	86.71% Global Rate
Undocumented	21	13.29%
Adoption by Creator		
Major Tech Actors	94.74%	High institutional resources
Independent	84.17%	Community-driven

There is more visual separation between the Independent trials than the other two. With the exception of a relatively small number of outliers, many of the Independent models were correctly identified as containing environmental disclosures. Hence, if the trial is actually an independent repository then it will probably be identified as such.

The Major Tech and Generative AI trials tend to be spread over an area centred around social bias and safety. At best, approximately 65.0% of these trials have been correctly populated. The distribution does not seem to have the kind of separation seen for the environmental trials, suggesting that they have difficulty documenting ecological costs, but could tell that social risks had varied.

4.1 Analysis and interpretation

The documentation adherence for different model domains is given in Table 2. They show similar measures for correctly identifying hazards in text-heavy domains; however, Computer Vision documentation exhibited lower adherence at 83.33%.

Table 2: Documentation Adherence by Model Domain (RQ4)

Domain	**Adoption Rate**	**Sample (n)**
Multimodal	100.00%	5
Audio	100.00%	2
Other	88.24%	17
NLP (Text)	86.36%	110
Computer Vision	83.33%	24

The mean scores recorded for all models are given in Table 3. They show similar measures for correctly identifying Bias and Safety: both have mean scores indicating high adherence in the Generative era, with the confusion

being predominantly between which of the two social risk controllers is operating. The Environmental Impact trials have a higher tendency to be omitted, with a score of 0.0% for corporate actors.

Each Major Tech model in the experiment had a higher score for identifying Bias than the other categories. It appears that Environmental Impact is the least identifiable of the three and the confusion tends to be between the social and safety sections. For analysis purposes, we can express the findings as polarized decisions. There is confusion between Bias and Safety, whereas the Environmental trials were identified only 8.9% of the time in the independent sample.

Table 3: Typology of Ethical Considerations (RQ2)

Theme	**% of Models**	**Description**
Bias & Fairness	65.0%	Gender, race, representation
Safety & Misuse	48.9%	Toxicity, jailbreaking, deepfakes
Privacy & PII	10.2%	Consent, surveillance
Environmental	6.6%	Carbon footprint, energy

4.2 Comparative tests

In order to test the distinguishability of one epoch or creator from the other, we performed a frequency analysis, calculated over all 158 models. Comparing the Pre-Generative Era and the Generative AI Era, we get the results shown in Table 4. A significant difference is found in the developers' identification of safety risks. Whilst the Generative AI era shares the characteristic of having variable documentation quality and thus is not identifiable simply by trying to detect a lack of README files, we would expect that if there was a "regression" characteristic to the GenAI response, then the auditor would be able to identify the era. It appeared that, generally, there was such a characteristic and adoption rates for safety increased substantially from 33.3% to 52.7%, while institutional environmental disclosures for Big Tech remained at 0.0%.

4.3 Ratings

In addition to the identification of the themes for each trial, we also asked each participant (auditor) to rate each model with respect to how well it had worked as a transparent system. The Generative AI models were consistently rated worse than the Pre-Generative ones. The differences between the Pre-Generative ratings and the others were found to be significant.

Table 4: Cross-Case Comparison: Temporal Shifts and Creator Priorities (RQ3 & Deep Dive)

Theme	**By Era (Temporal)**		**By Creator (Institutional)**	
	Pre-GenAI (2018-22)	**GenAI Era** (2023-25)	**Big Tech** (Corp)	**Independent** (Open Source)
Bias & Fairness	63.0%	**65.5%**	**77.8%**	60.4%
Safety & Misuse	33.3%	**52.7%**	**58.3%**	45.5%
Privacy & PII	7.4%	10.9%	16.7%	7.9%
Environmental	7.4%	6.4%	**0.0%**	**8.9%**

This analysis of documentation ratings is a relatively traditional evaluation, in line with Likert-scale approaches. Our results demonstrate that the framework of the empirical audit allows for such evaluation, but also adds the extra dimension of direct comparison with a human-centered standard (Digital Humanism). It is encouraging that not only did the independent developers generally receive a higher rating for environmental transparency, but that their performance was sufficiently distinct to confuse auditors as to which was a corporate mandate and which was an organic disclosure. This suggests that ethically the independent sector is performing its task well.

This analysis of documentation ratings is a relatively traditional evaluation, in line with Likert-scale approaches. Our results demonstrate that the framework of the empirical audit allows for such evaluation, but also adds the extra dimension of direct comparison with a human-centered standard (Digital Humanism). It is encouraging that not only did the independent developers generally receive a higher rating for environmental transparency, but that their performance was sufficiently distinct to confuse auditors as to which was a corporate mandate and which was an organic disclosure. This suggests that ethically the independent sector is performing its task well.

5 Discussion

The mixed-methods approach described in Section 3 was designed to evaluate the semantic quality of machine learning documentation, without reducing the complex sociotechnical landscape to binary compliance checks. In the analysis of the 158 models, we have seen some of the possibilities afforded by this method.

Firstly, the content analysis can extract a detailed reconstruction of the developers' conceptualisation of ethics. Our investigation of the "Greenwashing

Paradox" provides us with interesting detail on the interaction between such concepts as social fairness and environmental responsibility. We found that Model Cards provided by major corporations tend to include extensive information corresponding to social bias rather than ecological impact. These findings would be difficult to obtain by purely quantitative methods such as counting filled metadata fields.

However, we see evidence that the discourses obtained are influenced by the industrial context: the independent repositories, structured with less corporate oversight, produced discourse directly acknowledging carbon footprints; while the Big Tech actors, though more rigorous in formatting, produced restricted discourse with zero mention of environmental costs in our sample. It is clear that the design of documentation frameworks requires a careful balance: templates should be designed to encourage holistic transparency, while taking care not to lead developers into a "compliance checklist" mentality that prioritizes brand safety over invisible costs.

Secondly, the longitudinal analysis can produce a quantitative result on whether the industry is progressing towards the standards of Digital Humanism—despite the fact that we cannot evaluate such alignment directly. Our time-series study found that developers exhibited significant regression in documentation completeness between the Pre-Generative Era and the GenAI Era. Quantity ratings alone might suggest a robust ecosystem, but the semantic data tell us more: that the rapid expansion of Generative AI achieves its aim of rapid deployment at the cost of the meticulous documentation standards seen in earlier phases.

Our findings reflect two realities, corporate and independent; but further, they illustrate two philosophical attitudes. The major actors derive high adoption rates for social bias sections, but at the cost of imposing a predetermined, risk-averse conceptual framework onto the documentation, excluding environmental impact. Our analysis of the "Modality Bias" represents a fairly strong reactive attitude in which the key risks are determined by the media cycle (e.g., text toxicity) rather than the inherent risks of the technology (e.g., computer vision surveillance).

Having explored our findings, we are in a position to compare and contrast them with aspirations proposed by other investigators, and then to work towards recommendations on the applicability of different documentation strategies to different contexts.

5.1 Comparison with other approaches

A useful point of comparison is the original proposal due to Mitchell et al. (2019), involving nine distinct sections to ensure accountability. As previously discussed, this approach raises issues of selective adoption. Indeed, Bhat et al. (2023) investigate the documentation practice and find technical metadata to be present while qualitative ethical sections are treated as noise. Our findings extend this by revealing that even when ethical sections are present, they are "Ethically Cherry-Picked." The purely voluntary adherence method may therefore only be appropriate to cases in which the developer has intrinsic motivation to be transparent.

One advantage of the current voluntary method is that it is quick to administer. However, our study reveals that this results in a "Traceability Gap." A third alternative, which is worthy of further exploration, is to gather data via automated extraction. Analogies between the software engineering context (CI/CD pipelines) and the model training context suggest that methods could be adopted for automatically logging carbon metrics. However, there are some issues which would need to be addressed:

- Most importantly, the inference from compute hours to carbon impact requires more validation work before it can be relied on for standardized reporting.
- The evaluative role of "Brand Safety" also needs clarification: the high prevalence of Bias warnings suggests a focus on liability, but this is unlikely to be the whole story regarding model safety.

6 Recommendations and Future Work

From our studies, we suggest that an investigator or policy-maker wishing to formally evaluate or improve machine learning documentation should consider the following:

1. Is the documentation authored by a Major Tech Actor or an Independent? If the former, the auditor should be wary of "Greenwashing" and specifically look for omissions regarding environmental impact, as our data suggests a 0.0% voluntary reporting rate in this sector.
2. Is the model text-based or vision-based? If it is Computer Vision, we recommend a more rigorous scrutiny of the "Intended Use" sections, as our domain analysis shows these models lag behind in documentation adherence (83.33%) compared to NLP, despite their high potential for

surveillance harm.

3. Is the system released during a period of rapid expansion? The experimental design of governance must reflect market dynamics. We found that during rapid industry expansion, documentation quantity drops. Therefore, external audits are more critical during these periods than during "Pre-Generative" stability.

4. Can the ethical metrics be automated? Often the willingness of developers to report "invisible costs" will be fairly small, which raises issues for the statistical power of voluntary transparency. Approaches based on automated traceability (linking code directly to carbon trackers) may become viable for evaluating systems, although there are at present some issues to be resolved. We consider this a topic for future research, rather than an approach to be generally recommended at present, although we look forward to developments in this area.

One of the key themes in our recommendations is that the design of documentation policy should aim as far as possible to reflect the material reality of the model's production, not just its intended social use.

7 Conclusions

Traditional governance frameworks for artificial intelligence systems frequently lag behind rigorous standards for Digital Humanism. In this article we have considered approaches that may usefully be applied to evaluating these documentation practices, and have presented findings based on a mixed-methods analysis of 158 models.

Our analysis aims to characterise the conceptual structures developers bring to bear in rendering a Model Card in their corporate context. We found that the content describing ethical risks presents a selective coverage of sociotechnical aspects. Specifically, we revealed a "Greenwashing Paradox" where corporate actors prioritize social bias mitigation while systematically omitting ecological costs. Furthermore, we identified a "GenAI Regression," where the rush to deploy generative models has led to a prioritization of safety warnings over process transparency.

We hope that our recommendations (Section 6) may be a useful starting-point for others to conduct evaluations in authentic development contexts. More generally, we believe that this area is underexplored and needs much more research, such as the further development of automated approaches

to carbon tracking, or the application of rigorous third-party auditing to challenge the "negotiated truths" of corporate documentation.

References