Merging (EU)-Regulation and Model Reporting

Danilo Brajovic[1,3] **Vincent Philipp Göbels**[2] **Janika Kutz**[2] **Marco F. Huber**[1,3]
[1]Fraunhofer IPA [2] Fraunhofer IAO [3] Stuttgart University
danilo.brajovic@ipa.fraunhofer.de
{vincent-philipp.goebels, janika.kutz}@iao.fraunhofer.de
marco.huber@ieee.org

Abstract

Regulating AI systems remains a complex and unsolved issue despite years of active research. Various governmental approaches are currently underway, with the European AI Act [22] being a significant initiative in this domain. In the absence of official regulations, researchers and developers have been exploring their own methods to ensure the secure application of AI systems. One well-established practice is the usage and documentation of AI applications through data and model cards [28, 63, 54]. Although data and model cards do not explicitly address regulation, they are widely adopted in practice and share common characteristics with regulatory efforts. This paper presents an extended framework for reporting AI applications based on use-case, data, model, and deployment cards, specifically designed to address upcoming regulations by the European Union. The proposed framework aligns with industry practices and provides comprehensive guidance for regulatory compliance and transparent reporting. By documenting the development process and addressing key requirements, the framework aims to support the responsible and accountable deployment of AI systems in line with EU regulations, positioning developers well for future legal requirements.

1 Introduction

The development and regulation of AI systems pose unique challenges compared to traditional software development due to their stochastic nature. While policy makers have recognized this challenge and are actively working on regulatory approaches, these efforts are still in the early stages. Consequently, practitioners involved in AI application development often lack detailed knowledge about the exact regulatory requirements. This creates a difficult position for developers, as crucial steps may be inadvertently omitted during the development process, potentially rendering already developed applications non-compliant once regulations come into effect. To address this gap, this short paper proposes an extended framework for reporting AI applications based on data and model cards, specifically designed to address upcoming regulations by the European Union [22]. The full paper, currently under review, provides a more comprehensive exploration of the framework. By leveraging our experience working with large and medium-sized companies in Europe, we introduce a novel approach to report the development of AI applications based on four major development steps derived from prior work on data and model cards. Our primary contribution includes the introduction of use-case and deployment cards, as well as several updates to data and model cards, all aimed at meeting regulatory requirements.

1.1 Insights from Industry and Research Activities

To provide context for our work, we first present insights derived from various activities surrounding the safeguarding of AI systems and AI certification over the past two years. These activities include

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

collaborations on four projects with industry partners in Germany, spanning domains such as automotive, finance, and food manufacturing. Additionally, we conducted interviews with certification experts, employees, and developers to gather additional perspectives and insights.

The AI Assessment Catalogue as de facto standard

- The AI assessment catalog [62] is one of the first and most comprehensive works around safeguarding AI systems. It is highly regarded and used among industry players in Germany for safeguarding AI systems.
- While comprehensive, its length of approximately 160 pages made it challenging to use practically during development.
- The assessment catalog is designed for final assessments rather than development support, focusing on risk dimensions from human oversight to robustness.

Partners expressed uncertainty about the assessment catalog

- All our project partners were aware of the assessment catalog and partially utilized it in their development processes.
- Key concerns included compliance with the AI Act and the document's length.
- Despite these concerns, the assessment catalog served as the primary entry point for developing safe AI systems in German industry.

High hopes for standards

- Alongside the assessment catalog, our project partners were eagerly awaiting standards that would provide guidance on developing safe AI systems.
- Specific areas of interest included best practices for data collection and splitting, as well as selecting appropriate model classes.
- However, there were concerns about whether the forthcoming standards would offer the required level of detail.

Need for implementation details, technical tools, and specific numbers

- Our partners expressed a desire for more concrete guidance, seeking specific tools, methods, and algorithms to use in their AI development processes.
- Rather than general recommendations, they preferred actionable instructions and references to specific toolboxes or algorithms.

Trust-building as a motivation from industry

- Building trust in AI systems among end-users and affected individuals emerged as a key motivation for certification and assessment of AI applications.
- Interviews with individuals affected by AI systems indicated their desire to be included in the development process and understand design choices related to the AI system, which can be addressed through straightforward documentation.

Taking into account the feedback and insights mentioned above, our proposed framework adopts the format of cards, aligning with the common industry practice of reporting AI applications using data and model cards [28, 63, 54]. However, we introduce a different organizational approach by documenting the development of an AI application along the development process, in contrast to the organization along risk dimensions as used in the assessment catalog [62]. This approach allows for a more comprehensive reporting structure.

In addition to the existing data and model cards, our framework introduces the use-case and deployment cards. The use-case card provides a concise overview of the underlying problem and a risk classification based on established risk dimensions. The deployment card focuses on the reporting requirements for the deployment phase of AI applications, drawing primarily from the assessment catalog [62] and the AI Act [22].

Furthermore, our main contribution extends beyond the introduction of new cards. We include references to additional resources, toolboxes, and the regulatory sources, providing comprehensive guidance for regulatory compliance and transparent reporting. This integration of supplementary materials ensures developers have access to the necessary tools and references to meet regulatory requirements and support responsible AI deployment.

2 Proposed Framework

Before presenting our cards, we briefly cover similarities and differences of our proposed frameworks to previous work. However, we do not go into the technical details because it would exceed the scope of this paper.

2.1 Use-Case Card

In our proposed framework, we begin with a use-case card, which provides a concise description of the underlying problem and a risk classification based on established risk dimensions. While the concept of providing a general overview is not entirely new (e.g., Poretschkin et al. [62] utilizes a similar concept in the assessment catalog), our proposal differs by incorporating regulatory requirements and a risk classification. This inclusion benefits non-technical users and potential auditors by offering a quick overview of the use-case and solution approach. The early risk assessment ensures that developers adhere to the same risk dimensions throughout the development process. For instance, if fairness is identified as a risk, it will be addressed during data collection, model training, and deployment. Without this alignment, the responsible person for data collection may overlook fairness as a risk, leading to the omission of necessary data and subsequent developers being unable to mitigate potential discrimination.

2.2 Data Card

While data cards [63] and data sheets [28] are already established in industry and cover a wide range of comprehensive questions (typically exceeding 30), our proposed data card follows similar content themes while extending the requirements with additional recommendations for toolboxes and references to the regulatory sources from the AI Act [22]. These additional requirements aim to provide a more comprehensive framework for data documentation and address the specific needs of regulatory compliance.

It is worth noting that due to the scope of this short version, some requirements from our data card may initially appear unclear. However, these requirements are elaborated in more detail in the full paper, where we provide in-depth explanations, best practices, and considerations derived from recent research. For example, we discuss possible best practices for data splitting, which can contribute to the development of robust and fair AI models.

2.3 Model Card

Similar to data cards, model cards [54] have become a widely used framework in industry. However, our version differentiates by including a section on explainable AI due to the transparency requirement in the AI Act. Additionally, we have removed the data section from the original work, as it has been transferred to the data card.

2.4 Deployment Card

Our proposed framework introduces the deployment card, which focuses on the reporting requirements for the deployment phase of AI applications. To the best of our knowledge, no other scientific work specifically addresses a reporting framework for AI application deployment. As a result, our requirements draw primarily from the assessment catalog [62] and the AI Act [22], encompassing thirteen key steps ranging from AI application monitoring to testing and roll-out while leaning on key concepts and best practices of software development and DevOps. This tailored focus on deployment reporting sets our framework apart from existing approaches.

Use Case Card			
Step	**Requirement**	**AI Act**	**References**
General	Name contact person	Art. 13 (a)	
	List groups of people involved		[62]
	Summarize the use case shortly	Art. 11	[62]
	Describe the status quo		[62]
	Provide a short solution summary	Art. 11	[54]
	Review past incidents in similar use cases		AI Incident Database
	Review available tools		Tools&Metrics for Trustworthy AI
Problem definition	Describe the underlying technical challenges		[62]
	Formulate the learning problem	Art. 11	[60]
	Describe the disadvantages of the status quo		[69, 68]
	Argue why non AI approaches are not sufficient		[69, 68]
Solution approach	Describe the integration into the current workflow	Art. 11	
	Document the intended purpose of the AI	Art. 13 3. (b)	
	Document if a foundation model will be used or developed	Art. 3, Art. 28 b)	
	Provide a short data description	Art. 10	[28]
	Risk Assessment	Art. 9, Art. 19	
Risk Class	Categorize the application into a risk class according to the AI Act	Art. 5, Art. 6	AI Act Flow Chart; Risk Database
Human agency and oversight	Rate the level of autonomy	Art. 14	
Technical robustness and safety	Evaluate the danger to life and health	Art. 5 (1) a), Art. 14	[73]
	Identify possibilities of non-compliance	Art. 15, 19	[43]
	Identify and list customer relevant malfunction		[36, 61]
	Identify and list internal malfunctions and foreseeable misuse	Art. 13 3. (b) (iii)	[36, 61]
	Evaluate cybersecurity risks	Art. 15, 42	
Privacy and data governance	List risks connected with customer data		[17]
	List risks connected with employee data		[34, 79]
	List risks connected with company data		[14, 4]
Transparency	Evaluate effects of incomprehensible decisions or the use of a black box model	Art. 13	[48, 21, 23, 68, 13]
Diversity, non-discrimination and fairness	Check for possible manipulation of groups of people		[41, 5, 8]
	Check for discrimination of groups of people regarding sensitive attributes	Art. 5 (1) b) and c)	[41, 16, 59]
Societal and environmental well being	List dangers to the environment		[76]
	Consider ethical aspects		[19, 56, 7, 33, 15]
	Evaluate effects on corporate actions		[29]
	Identify impact on the staff		[51]
Accountability	Estimate the financial damage on failure		[31]
	Estimate the image damage on failure		[14, 4, 43]
Norms	List relevant norms in the context of the application (e.g., automotive safety norms)	Art. 9 (3)	[27]; AI Standards Hub
Involvement of Individuals	If feasible, describe involvement of affected individuals		

Data Card			
Step	**Requirement**	**AI Act**	**References**
General	Name originator of the dataset and provide a contact person		[63, 28]
	Describe the intended use of the dataset		[63, 28]
	Describe licensing and terms of usage		[63, 28]
Data-description	Collect requirements for the data before starting data collection	Art. 10 (2) d) e)	
	Describe a data point with its interpretation		[63, 28]
	Maybe, provide additional documentation to understand the data (e.g., links to scientific sources, preprocessing steps or other necessary information)		[63]
	If there is GDPR relevant data (i.e. personally identifiable information), describe it		[63, 28]
	If there is biometric data, describe it		[63, 28, 62]
	If there is copyrighted data, summarize it	Art. 28 b)	
	If there is business relevant information, describe it		[62]
Collection	Describe the data collection procedure and the data sources	Art. 10 (2) b) c), Annex III	[63, 28]
	Use data version control	Art. 10 (2) b)	[62]
	Consider the prior requirements for the data	Art. 10 (2) e)	
	Include and describe metadata		Metadata Standards
	Involve domain experts and describe their involvement		[62]
	Describe technical measures to ensure completeness of data	Art. 10 (2) g), Art. 28 b)	CleanLab
	Describe and record edge cases		[62, 63]
	If personal data is used, make sure and document that all individuals know they are part of the data		[50, 28]
	Describe if and how the data could be misused		[28, 63]
	If fairness is identified as a risk, list sensitive attributes	Art. 10 (5)	[41]
	If applicable, address data poisoning		[62]
Labeling	If applicable, describe the labeling process		[28, 63]; snorkel.ai
	If applicable, describe how the label quality is checked		
Splitting	Create and document meaningful splits with stratification	Art. 10 (3)	[6, 46, 65, 44, 81, 62]
	Describe how data leakage is prevented		[62, 73]
	Recommendation: test the splits and variance via cross-validation		[9]
	Recommendation: split dataset into difficult, trivial and moderate		[53]
	Recommendation: put special focus on label quality of test data		[58]
	Reminder: Perform separate data preprocessing on the splits		[67, 49]

Preprocessing	Document and motivate all processing steps that are a fixed part of the data	Art. 10 (2) c)	
	Document whether the raw data can be accessed		[28]
	If sensitive data is available highlight (pseudo)-anonymization	Art. 10 (5)	[62]
	If fairness is a risk, highlight fairness specific preprocessing	Art. 10 (5)	[62]
Analyzing	Understand and document characteristics of test and training data.	Art. 10 (2) e) g), Art 10 (3)	[37, 40, 81, 55, 52]
	Document why the data distribution fits the real conditions or why this is not necessary for the use case	Art. 10 (4)	[62]
	Document limitations such as errors, noise, bias or known confounders	Art. 10 (2) f) g)	[28, 63, 74, 62]
Serving	Describe how the dataset will be maintained in future		[28, 63, 50]
	Describe the storage concept (e.g., everything users need to know to access the data). For developers it must be possible to document on which version of the data a specific model was trained.		[62]
	Describe the backup procedure		[62]
	If necessary, document measures against data poisoning		[62]
Further notes	Document further recommendations or shortcomings in the data		
Involvement of Individuals	If feasible, describe involvement of affected individuals		

Model Card			
Step	**Requirement**	**AI Act**	**References**
General	Name model originator and provide a contact person		[54]
	Document the creation date and version of the model		[54]
	Describe intended use of the model	Art. 13 3. (b)	[54]
Description	Describe the architecture and size of the used model	Annex VIII	[54]
	Describe the used hyperparameters		
	Document the training, validation, and test error	Art. 13 3. (b) Art. 15	[54]
	Document computation complexity, training time and energy consumption (for foundation models describe steps taken to reduce energy consumption)	Art. 28 b), Annex VIII	[78, 26]
Explainability and interpretability	Document the required level of explainability and interpretability	Art. 13 (1, 2), Art. 15 (2)	[62, 35, 18]
	Describe taken actions if any	Art. 15 (2)	[35, 66, 57, 70, 11]
Feature engineering, feature selection and preprocessing	Describe whether interpretability was considered for feature engineering	Art. 13 (1, 2), Art. 15 (2)	
	Describe and justify (domain specific) feature engineering and selection	Art. 10 (3, 4)	[54, 45, 12]
	Describe and justify preprocessing steps		[54, 45]
Model selection	Document comparison to standard baselines, benchmarks and other evaluated models	Art. 28 b), Annex VIII	[49]; ML-Baselines; SOTA Models
	Justify the model choice and considerations regarding explanibility and interpretability	Art. 13 (1, 2), Art. 15 (2)	[49]
	Describe why the complexity of model is justified and needed		
	Document the approach of hyperparameter optimization		[9, 75, 77, 25, 30, 64]
	Describe the model evaluation	Art. 5, 6, 7, 9, Art. 28 b), Annex VIII	[81, 80, 10, 64, 42]
Choice of metrics	Describe selected metrics and justify them regarding use case and fairness	Art. 13 3. (b)	[54, 72, 24, 62, 41]
	Formulate the KPIs for go-live (domain specific reasons)		[62]
Model confidence	If relevant, document and quantify uncertainty of the model		[1]
	If relevant, document how uncertainty is handled		[32, 71]
Testing in real world setting	Describe the test design	Art. 5, 6, 7, 9	
	Describe possible risks, edge cases and worst case scenarios and create (or simulate) them if possible	Art. 15 (3), Art. 28 b), Annex VIII	
	Describe limitations and shortcomings of the model	Art. 28 b), Annex VIII	
	Describe the test results	Art. 28 b), Annex VIII	
	Explain the derived actions	Art. 28 b), Annex VIII	
More	Describe further recommendations or shortcomings	Annex VIII	
Involvement of Individuals	If feasible, describe involvement of affected individuals		

Operation Card			
Step	**Requirement**	**AI Act**	**References**
Scope and aim of monitoring	Describe monitored components	Art. 61 (1,3)	
	Assess risks and potential dangers according to Use Case Card	Art. 9 (2)	
	List safety measures for risks	Art. 9 (4)	[62]
Operating concept	Create and document utilisation concept	Art. 13, Art. 17, Art. 16	
	Plan staff training	Art. 9 (4) c)	[2]
	Determine responsibilities		[62]
Autonomy of application	Document decision-making power of AI	Art. 14, Art. 17	[62, 15]
	Determine process to overrule decisions of the AI	Art. 14 (4) d) e)	[62]
Responsibilities and measures	Document component wise: assessed risk, control interval, responsibility, measures for emergency	Art. 9	[62, 15, 2]
Model performance	Monitor input and output	Art. 17 (1) d), Art. 61(2)	[62]
	Detect drifts in input data		[62]
	Document metric in use to monitor model performance	Art. 15 (2)	
AI interface	Establish transparent decision making process	Art. 52	[62, 47]
	Establish insight in model performance on different levels	Art. 14	[62]
	Declare content created as product of the AI	Art. 52, Recital 60 g	
IT security	Document individual access to server rooms	Art. 15	[73, 3]
	Set and document needed clearance level for changes to AI/deployment/access regulation		[73, 3]
	Establish and audit ISMS		[38, 39]
Privacy	Justify and document use or waiver of a privacy preserving algorithm	Art. 10 (5)	[62, 20]
	If applicable, document privacy algorithm and due changes in the monitoring of output data		[20]
MLOps	Establish versioned code repository of AI, training and deployment		[62]
	Establish maintenance and update schedule		[62]
	Set regulation for the retraining of the AI and decision basis for the replacement of a model		[62]
Registration	For high-risk AI systems or foundation models, register the AI in the EU database	Art. 51, Art. 60, Annex VIII	[62]
Record keeping	Keep logs of events that include (at minimum) a time stamp, input data and allow identification of the person responsible for human oversight	Art. 12 4.	
Testing and rollout	Determine responsibilities for updates		[62]
	Document software tests	Art. 17 (1) d)	[62]
	Set period of time that an update must function stably before it is transferred to live status		[62]
Involvement of Individuals	If feasible, describe involvement of affected individuals		

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges, 2021.

[2] Martin Ahlfeld, Till Barleben, et al. Industrie 4.0 – how well the law is keeping pace, 2017. URL `https://www.plattform-i40.de/IP/Redaktion/EN/Downloads/Publikation/i40-how-law-is-keeping-pace.html`.

[3] Dennis Amschewitz, Gesmann-Nuissl, et al. Artificial intelligence and law in the context of industrie 4.0, 2019. URL `https://www.plattform-i40.de/IP/Redaktion/EN/Downloads/Publikation/AI-and-Law.html`.

[4] Arvind Ashta and Heinz Herrmann. Artificial intelligence and fintech: An overview of opportunities and risks for banking, investments, and microfinance. *Strategic Change*, 30(3):211–222, 2021.

[5] Hal Ashton and Matija Franklin. The problem of behaviour and preference manipulation in ai systems. In *CEUR Workshop Proceedings*, volume 3087. CEUR Workshop Proceedings, 2022.

[6] Claudia Beleites, Ute Neugebauer, Thomas Bocklitz, Christoph Krafft, and Jürgen Popp. Sample size planning for classification models, 1 2013. ISSN 00032670.

[7] Nick Bostrom and Eliezer Yudkowsky. The ethics of artificial intelligence. In *Artificial intelligence safety and security*, pages 57–69. Chapman and Hall/CRC, 2018.

[8] Johnny Botha and Heloise Pieterse. Fake news and deepfakes: A dangerous threat for 21st century information security. In *ICCWS 2020 15th International Conference on Cyber Warfare and Security. Academic Conferences and publishing limited*, page 57, 2020.

[9] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks, 2021.

[10] Houssem Ben Braiek and Foutse Khomh. On testing machine learning programs, 2018. URL `http://arxiv.org/abs/1812.02257`.

[11] Nadia Burkart and Marco F. Huber. A survey on the explainability of supervised machine learning, jan 2021. URL `https://doi.org/10.1613%2Fjair.1.12228`.

[12] Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. Feature selection in machine learning: A new perspective, 2018.

[13] Davide Castelvecchi. Can we open the black box of ai?, 2016.

[14] Benjamin Cheatham, Kia Javanmardian, and Hamid Samandari. Confronting the risks of artificial intelligence. *McKinsey Quarterly*, 2(38):1–9, 2019.

[15] European Commission, Content Directorate-General for Communications Networks, and Technology. Ethics guidelines for trustworthy ai, 2019.

[16] Council of European Union. Handbook on european non-discrimination law : 2018 edition, 2019.

[17] Emiliano De Cristofaro. An overview of privacy in machine learning, 2020. URL `https://arxiv.org/abs/2005.08679`.

[18] DIN SPEC 92001-3. Künstliche intelligenz - life cycle prozesse und qualitätsanforderungen - teil 3: Erklärbarkeit. Standard, DIN Deutsches Institut für Normung e. V., Berlin, DE, 2023.

[19] Markus Dirk Dubber, Frank Pasquale, and Sunit Das. *The Oxford handbook of ethics of AI*. Oxford Handbooks, 2020.

[20] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy, 2014.

[21] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2021.

[22] European Comission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021. ISSN 52021PC0206. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[23] Heike Felzmann, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns, 2019.

[24] César Ferri, José Hernández-Orallo, and R Modroiu. An experimental comparison of performance measures for classification, 2009.

[25] Matthias Feurer and Frank Hutter. Hyperparameter optimization, 2019.

[26] Eva García-Martín, Crefeda Faviola Rodrigues, Graham Riley, and Håkan Grahn. Estimation of energy consumption in machine learning, 2019.

[27] Urs Gasser and Carolyn Schmitt. The role of professional norms in the governance of artificial intelligence. In *The oxford handbook of ethics of AI*, page 141. Oxford University Press Oxford, 2020.

[28] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets, 2018. URL http://arxiv.org/abs/1803.09010.

[29] Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research, 2020.

[30] Varun Godbole, George E. Dahl, Justin Gilmer, Christopher J. Shallue, and Zachary Nado. Deep learning tuning playbook, 2023. URL http://github.com/google/tuning_playbook. Version 1.0.

[31] Francesco Gualdi and Antonio Cordella. Artificial intelligence and decision-making: The question of accountability, 2021.

[32] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[33] Thilo Hagendorff. The ethics of ai ethics: An evaluation of guidelines, 2020.

[34] Claretha Hughes, Lionel Robert, Kristin Frady, and Adam Arroyos. Artificial intelligence, employee engagement, fairness, and job outcomes. In *Managing Technology and Middle-and Low-skilled Employees: Advances for Economic Regeneration*, pages 61–68. Emerald Publishing Limited, 2019.

[35] ISO6254. Information technology — artificial intelligence — objectives and approaches for explainability of ml models and ai systems. Standard, International Organization for Standardization, Geneva, CH, 2023.

[36] ISO/IEC 23894. Information technology — artificial intelligence — guidance on risk management. Standard, International Organization for Standardization, Geneva, CH, 2023. URL https://www.iso.org/standard/77304.html?browse=tc.

[37] ISO/IEC 25012. Software product quality requirements and evaluation (square) — data quality model. Standard, International Organization for Standardization, Geneva, CH, 2008. URL https://www.iso.org/standard/35736.html.

[38] ISO/IEC 27000. Information technology — security techniques — information security management systems. Standard, International Organization for Standardization, Geneva, CH, 2018.

[39] ISO/IEC 27001. Information security management systems. Standard, International Organization for Standardization, Geneva, CH, 2022.

[40] ISO/IEC 5259x. Artificial intelligence — data quality for analytics and machine learning (ml). Standard, International Organization for Standardization, Geneva, CH, 2023. URL https://www.iso.org/standard/81088.html.

[41] ISO/IEC CD TS 12791. Information technology — artificial intelligence — treatment of unwanted bias in classification and regression machine learning tasks. Standard, International Organization for Standardization, Geneva, CH, 2023. URL https://www.iso.org/standard/83002.html.

[42] ISO/IEC TS 4213. Information technology — artificial intelligence — assessment of machine learning classification performance. Standard, International Organization for Standardization, Geneva, CH, 2022.

[43] Justin Johnson and D Daniel Sokol. Understanding ai collusion and compliance, 2020.

[44] V. Roshan Joseph. Optimal ratio for data splitting, 8 2022. ISSN 19321872.

[45] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in ml-based science, 2022. URL https://arxiv.org/abs/2207.07048.

[46] Frank Konietschke, Karima Schwab, and Markus Pauly. Small sample sizes: A big data problem in high-dimensional data analysis, 3 2021. ISSN 14770334.

[47] Tom Kraus, Lene Ganschow, Marlene Eisenträger, and Steffen Wischmann. Erklärbare KI: Anforderungen, Anwendungsfälle und Lösungen, 2021. URL https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/KI-Inno/2021/Studie_Erklaerbare_KI.pdf?__blob=publicationFile&v=1.

[48] Stefan Larsson and Fredrik Heintz. Transparency in artificial intelligence, 2020.

[49] Michael A. Lones. How to avoid machine learning pitfalls: a guide for academic researchers, 2021. URL https://arxiv.org/abs/2108.02497.

[50] Zhicong Lu, Rubaiat Habib Kazi, Li Yi Wei, Mira Dontcheva, and Karrie Karahalios. A framework for deprecating datasets: Standardizing documentation, identification, and communication, 4 2021. ISSN 25730142.

[51] Nishtha Malik, Shalini Nath Tripathi, Arpan Kumar Kar, and Shivam Gupta. Impact of artificial intelligence on employees working in industry 4.0 led organizations, 2022.

[52] Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Tariq Kane, Christine R. Kirkpatrick, Tzu-Sheng Kuo, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. Dataperf: Benchmarks for data-centric ai development, 7 2022. URL http://arxiv.org/abs/2207.10062.

[53] Kristof Meding, Luca M. Schulze Buschoff, Robert Geirhos, and Felix A. Wichmann. Trivial or impossible – dichotomous data difficulty masks model differences (on imagenet and beyond), 2021. URL http://arxiv.org/abs/2110.05922.

[54] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. pages 220–229. Association for Computing Machinery, Inc, 1 2019. ISBN 9781450361255. doi: 10.1145/3287560.3287596.

[55] Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. Measuring data, 12 2022. URL http://arxiv.org/abs/2212.05129.

[56] Vincent C Müller. Ethics of artificial intelligence and robotics, 2020.

[57] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Joerg Schloetterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, 2022.

[58] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks, 3 2021. URL http://arxiv.org/abs/2103.14749.

[59] Luca Oneto and Silvia Chiappa. Fairness in machine learning. In *Recent Trends in Learning From Data*, pages 155–196. Springer International Publishing, 2020. doi: 10.1007/978-3-030-43883-8_7. URL https://doi.org/10.1007%2F978-3-030-43883-8_7.

[60] Samir Passi and Solon Barocas. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 39–48, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287567. URL https://doi.org/10.1145/3287560.3287567.

[61] David Piorkowski, Michael Hind, and John Richards. Quantitative ai risk assessments: Opportunities and challenges, 2022. URL https://arxiv.org/abs/2209.06317.

[62] Maximilian Poretschkin, Anna Schmitz, Maram Akila, Linara Adilova, Daniel Becker, Armin B Cremers, Dirk Hecker, Sebastian Houben, Michael Mock, Julia Rosenzweig, et al. Leitfaden zur gestaltung vertrauenswürdiger künstlicher intelligenz (ki-prüfkatalog), 2021.

[63] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai, 4 2022. URL http://arxiv.org/abs/2204.01075.

[64] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning, 11 2018. URL http://arxiv.org/abs/1811.12808.

[65] S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners and open problems. volume 1, pages 417–423. Publ by IEEE, 1990. ISBN 0818620625. doi: 10.1109/icpr.1990.118138.

[66] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[67] David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, 2017. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.02881.

[68] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019.

[69] Cynthia Rudin and Joanna Radin. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition. *Harvard Data Science Review*, 1(2), nov 22 2019. https://hdsr.mitpress.mit.edu/pub/f9kuryi8.

[70] Nina Schaaf and Philipp Wiedenroth, Saskia Johanna ans Wagner. Explainable ai in practice - application-based evaluation of xai methods, 2022.

[71] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction., 2008.

[72] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Bias preservation in machine learning: The legality of fairness metrics under eu non-discrimination law, 2021. URL https://ssrn.com/abstract=3792772.

[73] Philip Matthias Winter, Sebastian Eder, Johannes Weissenböck, Christoph Schwald, Thomas Doms, Tom Vogt, Sepp Hochreiter, and Bernhard Nessler. Trusted artificial intelligence: Towards certification of machine learning applications, 2021.

[74] Nicole Wittenbrink, Tom Kraus, Stefanie Demirci, and Sebastian Straub. Leitfaden fÜr das qualitÄts-management bei der entwicklung von ki-produkten und services, 12 2022. URL https://www.digitale-technologien.de/DT/Redaktion/DE/Kurzmeldungen/Aktuelles/2022/KI-Inno/20221220_Leitfaden_Qualitaetsmanagement.html.

[75] Tzu-Tsung Wong and Po-Yang Yeh. Reliable accuracy estimates from k-fold cross validation, 2019.

[76] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. Sustainable ai: Environmental implications, challenges and opportunities, 2022.

[77] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice, 2020. URL https://arxiv.org/abs/2007.15745.

[78] Tien-Ju Yang, Yu-Hsin Chen, Joel Emer, and Vivienne Sze. A method to estimate the energy consumption of deep neural networks. In *2017 51st asilomar conference on signals, systems, and computers*, pages 1916–1920. IEEE, 2017.

[79] Serap Zel and Elif Kongar. Transforming digital employee experience with artificial intelligence. In *2020 IEEE / ITU International Conference on Artificial Intelligence for Good (AI4G)*, pages 176–179, 2020. doi: 10.1109/AI4G50087.2020.9311088.

[80] Jie M Zhang, Mark Harman, Lei Ma, and Yang Liu. Machine learning testing: Survey, landscapes and horizons, 2020.

[81] Xiaoyu Zhang, Jorge Piazentin Ono, Huan Song, Liang Gou, Kwan Liu Ma, and Liu Ren. Sliceteller: A data slice-driven approach for machine learning model validation, 2022. ISSN 19410506.