A Multidimensional Approach to Ethical AI Auditing

Sónia Teixeira[1], Atia Cortés[2], Dilhan Thilakarathne[3], Gianmarco Gori[4] Marco Minici[5], Monowar Bhuyan[6], Nina Khairova[6], Tosin Adewumi[7] Devvjiit Bhuyan[6], Jack O'Keefe[8], Carmela Comito[5], João Gama[1] and Virginia Dignum[6]

[1]INESC TEC
[2]Barcelona Supercomputing Center
[3]ING
[4]Vrije Universiteit Brussel
[5]ICAR-CNR
[6]University Umeå
[7]University Luleå
[8]Northwestern University

Abstract

The increasing integration of Artificial Intelligence (AI) across various sectors of society raises complex ethical challenges requiring systematic and scalable oversight mechanisms. While tools such as AIF360 and Aequitas address specific dimensions, namely fairness, there remains a lack of comprehensive frameworks capable of auditing multiple ethical principles simultaneously. This paper introduces a multidimensional AI auditing tool designed to evaluate systems across key dimensions: fairness, explainability, robustness, transparency, bias, sustainability, and legal compliance. Unlike existing tools, our framework enables simultaneous assessment of these dimensions, supporting more holistic and accountable AI deployment. We demonstrate the tool's applicability through use cases and discuss its implications for building trust and aligning AI development with fundamental ethical standards.

Introduction

The growing adoption of Artificial Intelligence (AI) across various sectors, including government and private companies, has revealed complex ethical challenges. Ensuring responsible development and deployment of AI systems requires addressing critical ethical considerations such as preventing bias and discrimination, ensuring fairness, respecting human autonomy, promoting social and environmental well-being, maintaining transparency and accountability, and protecting against harm (Cortes, Cortes, and Barrue 2019). In Europe, these principles are fundamental to creating trustworthy AI aligned with ethical, legal, robust (AI HLEG 2019) and societal standards. Similarly, UNESCO's *Recommendation on the Ethics of Artificial Intelligence* highlights fundamental principles, including human dignity, harm prevention, justice, transparency, responsibility, privacy, and sustainability, as guidelines to promote AI's positive impact on society (UNESCO 2022). Despite the growing adoption of AI, the consistent application of ethical principles and requirements such as fairness, explainability, transparency, and robustness remains an urgent issue. These principles are widely recognised by several organisations, including the *Assessment List for Trustworthy Artificial Intelligence* (ALTAI) (Independent High-Level Expert Group on Artificial Intelligence 2019). Many of these principles are shared across different recommendations, reflecting their fundamental nature and their connection to legal and societal values, such as non-discrimination. For example, AI RMF and the OECD AI Principles as away of ensuring that the development and use of AI is trustworthy, human-centered, and respects fundamental rights.

In this context, the European Union AI Act (AI Act 2024) marks a significant step in transforming the principles developed throughout AI practice and AI ethics into enforceable legal requirements. In particular, AI systems that are classified as high-risk under the AI Act must comply with requirements that include data governance, accuracy, robustness, transparency, and human oversight. The German Credit case study carried out in our work essentially focuses on the topic of AI-based credit decision making, a use-case classified as high-risk under Annex III, 5(b) of the AI Act, thereby contributing to a better understanding of the role that AI audit and evaluation play in meeting not only ethical but also legal requirements. The RUWA case study, focused on bias and trustworthiness in news reporting, contributes to the discussion on the assessment of the risk of dissemination and amplification of illegal content and online disinformation, which represents one of the key obligations of Very Large Online Platforms and Very Large Online Search Engines under Articles 34 and 35 of the EU Digital Services Act (of the European Parliament and of the Council of 19 October 2022 2022). In addition to regulatory frameworks like the AI Act, international standards such as ISO/IEC 25059:2023 and ISO/IEC 23894:2023 emphasise the importance of evaluating and auditing AI systems. These standards focus on ensuring software quality, mitigating societal and ethical risks, and promoting principles such as accountability, fairness, transparency, explainability, and respect for the rule of law. Existing tools such as AIF360 (Bellamy et al. 2018) and Aequitas (Saleiro et al. 2018) assess fairness but lack a com-

prehensive evaluation of multiple ethical principles simultaneously. To address this gap, this article presents an integrated tool to evaluate the multidimensions of AI ethics. The tool enables simultaneous analysis of key principles, including fairness, explainability, robustness, transparency (in data sets and AI systems), bias, sustainability, and legal aspects, as part of a broader trustworthiness assessment, making it a valuable instrument to promote AI literacy (Article 4 of the AI Act).

This work primarily focuses on the key principles for an ethical AI evaluation and their alignment with EU legislation. Specifically, it highlights the importance of analyzing multiple principles, how to assess them, and how to present them, an approach we realize through case studies. The dashboard is a way to present all these principles, offering an integrated perspective on their presentation (as existing approaches have so far addressed them separately). The main contributions of this work are: (i) the proposal of a novel multidimensional auditing framework that operationalises key ethical principles aligned with the EU AI Act, (ii) the development and implementation of an interactive tool for ethical auditing across datasets and models, and (iii) the application of this framework in two diverse case studies (credit and media), demonstrating applicability to real-world and high-risk AI contexts, (iv) a critical reflection on trade-offs between fairness, bias, explainability, and robustness, transparency, promoting ethical awareness in AI design. Our approach emphasises the evaluation of critical dimensions of ethical and reliable AI systems, such as data quality, bias, the explanations provided, accuracy, robustness, and outcomes (e.g. fairness). We used two data sets, RUWA and German Credit data sets, and applied different methods and techniques. Additionally, the transparency evaluation process, as well as the trustworthy evaluation process (for example), are cross-cutting, and the questionnaires that form the basis of the evaluation are not limited to the presented case studies, although they also serve as a way to assess these cases. Overall, this research contributes to understanding the complex interplay between normative, ethical and legal requirements, on the one hand, and the technical specifications and concrete design and development choices adopted to implement them, on the other hand (Gori 2024a).

Foundations of Ethical and Trustworthy AI

The following subsections emphasise the essential concepts for ethical development and provide background for the core dimensions of this work.

Source of Bias

One of the primary challenges in developing ethical AI systems lies in the issue of bias, which can stem from two key sources: biased datasets and biased algorithms (Schwartz et al. 2022). AI models, including machine learning (ML), deep learning (DL), and generative AI, utilize large datasets for training. If these data sets contain inherent biases—such as over- or under-representation of certain groups, perspectives, or scenarios—the AI model will learn and perpetuate those biases, according (Ntoutsi et al. 2020). For example, if a dataset used to train a facial recognition system includes fewer images of certain ethnic groups, the model may perform poorly on those groups, leading to biased outcomes (Leslie 2020). Similarly, if generative AI systems or large language models (LLMs) are trained on datasets that predominantly contain text from English-speaking sources, particularly from Western countries, the generated texts may overrepresent Western cultural norms, idioms, and perspectives (Liyanage and Ranaweera 2023). Bias in a dataset can arise from a variety of factors. Traditionally, one reason is the lack of a representative sample of the population, texts, or situation it is intended to generalize. As a case in point, the training of large language models predominantly on English-language texts from Western sources can be attributed to the limited availability of extensive language resources for the majority of the world's 7,000 languages (Le Scao et al. 2022), (Liu et al. 2024). An additional contributor to bias in training datasets may be labelling bias, also referred to as annotation bias. This occurs when different annotators assign inconsistent labels to similar data instances. For example, in a sentiment analysis dataset (Shah and Sureja 2024), (Ibrohim, Bosco, and Basile 2023), if one annotator categorizes a neutral text as positive while another categorizes it as negative, the model may learn contradictory patterns, undermining its ability to generalize effectively. A more complex scenario arises in datasets annotated for true/-false classification, such as those used in fact-checking or misinformation detection tasks. In these cases, the annotations may be influenced by the political, cultural biases or sector-specific contexts of the annotators, according to (Rangapur, Wang, and Shu 2023), which can lead to systematic errors in labelling. Such biases can significantly affect the reliability and validity of the dataset, ultimately impacting the performance and fairness of misinformation detection models trained on this data. However, even when datasets are balanced, algorithms themselves can introduce bias due to the way they process and prioritise information. For example, certain algorithmic architectures or design choices may favour specific patterns or correlations in the data, which can result in biased decision-making (Pagano et al. 2023).

User-Centered Explanations

Explainability in AI refers to the extent to which the processes and outcomes of an AI system can be understood by humans, a principle of particular importance in Europe, where ethical guidelines emphasize transparency and accountability in AI development. Traditional methods of explainability often rely on post-hoc analyses, where developers attempt to interpret AI decisions retrospectively (Samek, Wiegand, and Müller 2017), (Guidotti et al. 2018). Although these approaches offer insights, they frequently fail to address the diverse concerns and comprehension levels of users (Gilpin et al. 2018) and may not adequately reveal or mitigate the biases embedded in AI algorithms, potentially perpetuating unfair outcomes (Barocas, Hardt, and Narayanan 2019). A participatory approach ensures that explanations are not only technically robust but also meaningful to end-users. This method helps identify usability issues early, aligns system functionality with user needs, and fosters trust

and acceptance (Spinuzzi 2005). By directly involving users in the design process, participatory methods increase satisfaction and system reliability, as users see their contributions reflected in the outcomes (Schuler and Namioka 1993). Explainable AI presents two main challenges: developing techniques to extract explanations from black-box models and designing effective user interfaces for presenting these explanations (Panigutti et al. 2023). Despite the relatively recent focus on this area, several studies have applied participatory design methodologies. (Mucha et al. 2020) introduced a co-design approach to examine how users perceive and interact with explanations in AI decision support systems. Similarly, (Panigutti et al. 2023) adopted iterative prototyping and testing to design explainable AI techniques and interfaces for clinical decision support. Participatory methods have also been applied to design interfaces that calibrate trust in human-AI collaboration (Naiseh et al. 2024). Furthermore, (Bobek et al. 2024) explored user-centered perspectives on how people understand and interact with algorithmic explanations in the context of mushroom identification.

Fairness

Fairness and associated ethical principles are increasingly becoming essential to enhance the trustworthiness of AI systems. However, achieving fairness is complex, requiring a careful balance of various contexts, sociocultural factors (beyond just race, gender, and age), and ethical considerations, such as whether the implications of AI are appropriate for society. Among many factors, the data is often found to be a significant source of bias, with issues such as bias in collecting and sampling data affecting any models developed from it (Dablain, Krawczyk, and Chawla 2024). Moreover, the misuse of data leads to conscious or unconscious biases that result in developed biased models and unexpected decisions. Fairness in machine learning addresses either the technical dimensions of bias and fairness or delves into the theoretical discussions surrounding the social, legal, and ethical implications of discrimination in AI. Strategies for addressing fairness are typically employed at various stages of the machine learning pipeline: pre-processing, in-processing, and post-processing (Caton and Haas 2024). These approaches mainly focus on implementing interventions to mitigate bias. However, evaluating fairness lacks (Dablain, Krawczyk, and Chawla 2024) when combined with other ethical principles.

Robustness

We here try to provide a thorough overview of the various definitions of robustness and examines case studies that have addressed the issue of robustness in ML systems. *Adversarial Robustness.* Adversarial robustness is defined as a model's ability to maintain performance when confronted with adversarial inputs—carefully crafted perturbations designed to deceive the model. (Szegedy et al. 2013) first highlighted this vulnerability in neural networks, demonstrating that even imperceptible perturbations can lead to incorrect predictions. Subsequent work by (Goodfellow, Shlens, and Szegedy 2014) proposed the use of adversarial training as a mechanism to enhance model robustness against such attacks. *Distributional Robustness.* Distributional robustness refers to a model's performance under changes in the input data distribution. While distributional robustness was initially discussed in the context of linear programming by (Ben-Tal and Nemirovski 1999), it has since been adapted to ML systems. Recent work by (Sinha, Namkoong, and Duchi 2018) has advanced the understanding of distributional robustness in ML through principled adversarial training techniques. *Generalization Robustness.* Generalization robustness concerns a model's ability to generalize from the training data to unseen data. (Zhang et al. 2017) discussed this in terms of overfitting, where a robust model is one that avoids fitting noise in the training data and instead captures the underlying patterns that generalize well to new data. *Robustness to noise.* Models often encounter noisy or incomplete data in practical applications. Huber's foundational work in the 1960s, (Huber 1965), provided a formal framework for robust statistics, which has since been adapted to the ML context to enhance models' resilience against noise (Bishop 1995; Szegedy et al. 2013; Goodfellow, Shlens, and Szegedy 2014; Hendrycks and Gimpel 2016; Song et al. 2022). While there are numerous case studies on the robustness of ML systems across various domains, such as (Chen et al. 2015; Bertsimas, Pawlowski, and Zhuo 2018; Finlayson et al. 2019), significant gaps remain. Notably, there is a lack of case studies focused on the robustness of high-risk systems, such as credit scoring, against adversarial manipulations. Additionally, the field lacks a standardized metric for comparing the robustness of different ML models (Palumbo, Carneiro, and Alves 2024), making it challenging to evaluate and benchmark their resilience across diverse applications.

Transparency

The development of some of the most trending AI innovations over the last decade has often been controversial due to the appearance of significant failures related to biases that led to discrimination against social or ethnic groups. Some of these risks could have been avoided, or at least mitigated, with a proper mechanism of traceability and documentation prior to the deployment. Transparency is one of the seven key ethical requirements defined by the High-Level Expert Group on AI in the Ethics Guidelines for Trustworthy AI, (AI HLEG 2019). It has become a field of interest to AI researchers and industry and is now taking even more relevance due to the approval of the AI Act, which has a specific article requiring technical documentation that will contribute to enhancing the transparency of AI-based systems and, hence, trustworthiness. It refers to the extent to which an AI-based system can be understood by humans. This includes the following components:

- *Decision-making process:* to develop explainable and interpretable methods that will allow stakeholders to understand how the decisions of the AI-based system were.
- *Data transparency:* to provide documentation regarding data source, collection and usage for training and validation, as well as means to inform users about how data is

being used (in particular if it refers to personal data) or if there are any particular limitations related to the data.

- *Algorithmic transparency:* to describe the system's architecture, the design specifications, governance mechanisms, etc.

Transparency is a fundamental component for the ethical and responsible development of AI technologies, enabling other ethical requirements such as accountability, fairness and societal well-being. Technical documentation is essential and will soon become mandatory for high-risk AI applications. There are already several initiatives that have become popular as reporting tools, checklists such as *Datasheets for Datasets*, (Gebru et al. 2021), or *Model Cards for Model Reporting*, (Mitchell et al. 2019). However, these solutions present also some limitations. For instance there is yet no standard form of reporting and it is still uncertain what will be mandatory to be compliant with the AI Act. Also, these tools are not general enough to cover all the requirements from the AI Act (Hupont Torres et al. 2023), but at the same time they are not sensitive to the context or sector in which the AI application is developed.

Trustworthiness Assessment

The Canadian government's *Algorithmic Impact Assessment Tool* (AIA) and Microsoft's *Responsible AI Dashboard* (RAD) are two related works that serve different purposes than the *AI Guidelines Questionnaire for Companies* (AGQC), (Microsoft Corporation 2021), (RAIN Group 2021) and (Government of Canada 2020). The AIA is closely related to the AGQC in that it provides a numerical score outlining a project's adherence to Canada's AI safety procedures, similar to how the AGQC provides a project's risk score for a more general AI safety standard. It differs in that it additionally requires essay questions and does not visualize the results at the end. The safety categories are different as well with the AIA having six risk areas and two mitigation areas as opposed to 15 general categories in the AGQC. The RAD and AGQC share a common goal of assisting teams in their creation of responsible AI applications. The RAD is a more general tool than the AGQC as it uses Microsoft Copilot to give real-time recommendations and analysis of a project's progress. It does not provide a score related to a crystallized set of categories, but does provide an analysis of disparities in accuracy for different demographics and other failure modes of responsible AI. It also allows for content filters and other methods of algorithm tuning to ensure a project is rolled out safely.

Case Studies and Multidimensional Approach

This section provides an overview of the datasets, experimental setup, methodology to audit the key ethical dimensions of the dashboard, and the resulting findings. In general, for most of the principles considered, the approaches exhibit some degree of innovation (e.g., evaluation metric or methodological approach). All experiments using the *German Credit* dataset use at least a Decision Tree (DT) classifier in the latter.

RUWA Case

In this subsection, we present the RUWA case.

Data set The Russian-Ukraine War dataset (RUWA dataset), (Khairova et al. 2024), is a novel collection comprising over 16,000 news articles in English from global outlets covering the events of the Russian-Ukraine war. It includes articles from Europe, the USA, Ukraine, and Russia, covering the period from February 2022 to September 2022. Furthermore, all the news articles in the dataset were categorized into nine well-known events (or topics).

Evaluating the Truthfulness of News Texts To assess the truthfulness of news articles, given the insufficient number of annotated data available today, (Mishra, Shukla, and Agarwal 2022), we consider an unsupervised machine learning approach. As the primary dimension of our analysis, we investigate textual similarity by comparing news articles to determine their semantic congruence. Our objective is to ascertain whether news coverage of the same event from various global outlets exhibits sufficient similarity to suggest mutual reflection, or if they diverge significantly, which may indicate that some sources lack trustworthiness. We propose that news shared by media outlets in the two nations directly involved in the conflict will likely exhibit considerable differences. Therefore, the semantic similarity coefficient between texts from Russian and Ukrainian outlets will be minimal. Furthermore, We hypothesize that if articles covering the same event from most outlets show high semantic similarity, but one or two specific websites significantly diverge, this may indicate the potential untrustworthiness or untruthfulness of those websites. Methodologically, we employ a pairwise evaluation of the semantic similarity among articles from various outlets, aggregating all articles from the same source into a single textual document and utilizing FastText, Mikolov et al. (2017), for encoding. The experiments conducted provide robust validation of our hypotheses. Specifically, an analysis of news articles from outlets representing the two countries involved in the conflict, including *Censornet, Ukrinform, News Front*, and *RT*, reveals significant disparities across most reported events. These differences are systematically reflected in the semantic similarity coefficients, highlighting the distinct reporting styles and perspectives adopted by these outlets in the context of the ongoing conflict. The disparities in semantic similarity coefficients serve as an important indicator of potential bias and trustworthiness in news reporting.

Bias or misinformation Therefore, we may infer with a certain degree of confidence that the semantic similarity coefficient's value correlates with the likelihood of conveying a certain degree of consistency in narrative framing and factual alignment between the news sources. Higher similarity coefficients suggest that the outlets likely report similar information or adopt parallel perspectives, whereas lower coefficients may indicate divergent or conflicting accounts, possibly reflecting bias or misinformation. However, there is a difference between biased news coverage and misinformation, and semantic similarity metrics can reveal this distinction. While high semantic similarity coefficients typ-

ically indicate that outlets are reporting similar information, this alone does not necessarily imply that the information is unbiased or accurate. Bias can be detected when outlets report similar facts but frame them differently based on political or ideological perspectives. On the other hand, misinformation is typically reflected in low semantic similarity coefficients. Outlets spreading misinformation often present distorted or entirely fabricated accounts, diverging significantly from more reputable global sources. When an outlet's articles have a low similarity to others covering the same event, it could signal that the information is not just biased but potentially inaccurate or deliberately false.

German Credit Case

In this subsection, we present all the dimensions and approaches for the German Credit case (Figure 1a).

Data set The German Credit dataset (GCD), (Hofmann 1994), is commonly used in machine learning classification tasks. It contains 1000 observations and 21 variables, providing information about credit applications, such as applicants' characteristics and whether those applications were granted. The data set features a variety of financial, personal, and employment information from credit applicants, including age, gender, marital status, credit history, loan amount, loan duration, and more. Additionally, each entry in the dataset is labelled with the credit decision ("good" or "bad").

Sensitive Attribute Age and gender are two sensitive attributes that may be used to categorize people as privileged or unprivileged groups. In this work, we used the age attribute with a cut-off of 25, following, Kamiran and Calders (2009), where those equal to or older than 25 are the privileged group while the others are unprivileged.

Audit Bias We evaluate the training data for bias after splitting the whole dataset in the ratio 70:15:15. We used the *AI Fairness 360* (AIF360) library, by (Bellamy et al. 2019), which is a library of many state-of-the-art (SotA) algorithms on bias and fairness estimations and mitigation strategies. We used 7 metrics and their explanations to probe the dataset. These are Disparate Impact Ratio (DIR) by (Feldman et al. 2015), Mean Difference (MD) by (Thissen, Steinberg, and Gerrard 1986), Consistency by (Zemel et al. 2013), Number of Positives (NoP), Number of Negatives (NoN), Smoothed Empirical Differential Fairness (SEDF) by (Foulds et al. 2020), and Base rate by (Cao and Banaji 2016). Furthermore, we trained the decision trees (DT) classifier and evaluated the validation set. We also trained the logistic regression and evaluated the validation set based on accuracy and area under the curve (AUC). We performed a hyperparameter search over L1 and L2 regularization during training to determine the best hyperparameters. L2 regularization provided the best result. We have also used the *Fairlearn*, (Bird et al. 2020), package to incorporate two crucial bias mitigation techniques, namely the *ExponentiatedGradient*, and the *Threshold Optimizer*. The results related to the application of Fairlearn are discussed in this section. The validation accuracy and F1 score, using DT, are 0.707 and 0.71 while the accuracy and AUC results, using logistic regression, are 0.72 and 0.7585, respectively. The DIR result of about 0.8338 means the privileged group has favourable outcomes. If it had been greater than 1 then the unprivileged group would be deemed as having favourable outcomes. 1 would have meant no disparate impact. The MD result of -0.1185 implies the unprivileged group has less favourable outcomes, i.e. the privileged group gets 11.85% more positive outcomes.

Audit Explainability AI models are becoming increasingly complex and less transparent, highlighting the need to explain their decisions. The explanation of credit rejection decisions employs participatory design. The *Explainability* approach consists of two parts: the questionnaire and the clustering of features. Initially, a questionnaire was applied to identify the most effective method of communicating the reasons behind the rejections. In the next step, we applied the method considered most understandable by the questionnaire participants to all rejection cases, with the aim of visualizing which groups of features contributed most to the rejection.

- *Method selection:* The survey, conducted in English, consisted of two single-choice questions: one to identify which method generates the most understandable explanations and another to evaluate opinions on which method provides the most acceptable explanations for credit rejection. Participants were instructed to place themselves in the position of credit applicants and analyzed three rejection cases with explanations provided by different methods. Due to their widespread use in the literature, the chosen methods were LIME (Ribeiro, Singh, and Guestrin 2016), Shapley (Sundararajan and Najmi 2020), and Counterfactuals (Guidotti 2022). To ensure the survey's feasibility and proper understanding, it was pre-tested with three AI researchers, all native Portuguese speakers. The questionnaire responses were collected in an academic context from 36 individuals in March 2024.
- *Feature Analysis:* After identifying the most understandable method, we analyzed the features influencing credit rejection. We considered all cases where the prediction rejected the credit despite the data indicating approval. We used the *K-means* algorithm from the *factoextra* package (Kassambara and Mundt 2020), in R software (Team 2020), to cluster the data and generate a heatmap (Figure 1a) showing the importance of the features in the rejection.

Observations: the approach to *Explainability* require two steps: the questionnaire and the analysis of features. The questionnaire results indicated that the *Shapley* method generated the most understandable and realistic explanations for credit rejection. To understand which features contributed the most to the rejection, we analyzed the cases where the model rejected the credit. The analysis revealed that features such as *Phone, Purpose, Duration, Employment*, and *Account_status* were the most frequent in the rejection explanations. Additionally, patterns of behaviour between features, such as *Other_plans* and *Resident_since*, were found to be influential, although not always the most significant

(Figure 1a). The presence of *Foreign* status was not a determining factor for rejection, unlike *Gender* or *Age*.

Limitations: The adopted methodology, with participatory design, allowed for obtaining meaningful explanations for the user. However, if other methods for generating explanations (e.g., Anchors) were analyzed, the results might differ. Therefore, these are not the only possible explanations, but under the conditions mentioned above, they align with the most realistic explanations for credit rejection from the user's perspective.

Audit Fairness We focus on the sensitive attributes (SA) *Age* and *Gender* in the GCD dataset for our experiments. The *Age* attribute has an imbalance ratio of 5.71, while *Gender* has a ratio of 2.22. Key metrics, including *False Positive Rate* (FPR), *False Negative Rate* (FNR), and *Selection Rate* are used to evaluate fairness across demographic groups. On top of 'classical' metrics, we employed *Accuracy Score Difference, Demographic Parity Difference, Equalized Odds Difference, False Negative Rate Ratio*, and *False Positive Rate Ratio* for carried experiments. These metrics help identify discrimination among groups and mitigate biases in societal inequalities.

1. *Bias Measure*: This metric is used to determine the improvement in model fairness. It is a composite metric, formula (1), which is a non-linear combination of all the metrics presented in Table 1. This will have a range between [0, 1], where 1 represents the ideal model. The bias measure is computed as:

$$\begin{aligned} \text{bias\_measure} = {} & 0.32 \times \text{accuracy\_score\_diff} + \\ & 0.32 \times \text{demographic\_parity\_diff} + \\ & 0.12 \times \text{equalized\_odds\_diff} + \\ & 0.12 \times \text{sigmoid}(\text{FNR\_ratio}) + \\ & 0.12 \times \text{sigmoid}(\text{FPR\_ratio}) \end{aligned} \tag{1}$$

Among many methods available in Fairlearn (Bird et al. 2020), we chose to employ the following methods for bias mitigation according to the context considered in our experiments.

1. *ExponentiatedGradient*: A fairness-aware optimization algorithm adjusts classifier parameters or weights to minimize fairness violations while preserving accuracy. Using exponentiated gradient descent, it updates weights exponentially based on the fairness constraint gradient to balance accuracy and fairness.
2. *ThresholdOptimizer*: By adjusting a classifier's decision thresholds, this method enhances fairness while balancing accuracy. It employs convex optimization to refine thresholds over a loss function that integrates fairness constraints, reducing group disparities while maintaining predictive performance.

Our experiments begin with the use of *ExponentiatedGradient* over a range of constraints, determining the perfect constraint, and then running *ThresholdOptimizer* over the models to obtain the fair model that provides possible societal equalities. Various bias mitigation methods such as *Unmitigated, Demographic Parity, Equalized Odds, TPR Parity, FPR Parity* and *Threshold Optimizer* have been applied. The 'Unmitigated' model, optimized solely for accuracy, serves as a baseline and is expected to exhibit the highest bias. This model was subsequently optimized for fairness. Among the methods, the Demographic Parity (DP) constraint achieved the most significant reductions in both Demographic Parity Difference (DPD) and Equalized Odds Difference (EOD) for the *Age* attribute, with a stronger impact on DPD. However, since DP focuses only on equalizing positive outcomes across groups, it was less effective in addressing false positives, leading to a modest EOD reduction. For the *Gender* attribute, the Equalized Odds (EO) constraint performed better by minimizing both False Positive Rate (FPR) and False Negative Rate (FNR) disparities. With a lower imbalance ratio for *Gender*, EO effectively mitigated bias, while DP struggled, likely due to overcompensating selection rates. EO was less effective for *Age* due to the high imbalance ratio (5.71), where attempts to reduce FPR and TPR created unintended fairness trade-offs. Additionally, the Decision Tree model exhibited inherently lower bias for *Gender* than for *Age*, further influencing these outcomes. The Threshold Optimizer, which adjusts decision thresholds, was less effective in mitigating bias in the Decision Tree model. While decision trees inherently lack thresholds, sklearn's implementation allows probability outputs if the tree is sufficiently deep. The optimizer adjusted these probabilities but yielded limited improvements in fairness. All experimental results are presented in Table 1, for details.

Audit Robustness We propose an auditing framework to rigorously assess the robustness of high-risk AI systems under realistic and stealthy manipulation attacks as mandated by Article 15 of the EU AI Act. We fill a significant gap in the field, as, to the best of our knowledge, no comprehensive method currently exists for this purpose (Palumbo, Carneiro, and Alves 2024). We aim to test the capability of an high-risk AI system to consistently provide reliable decisions, even when inputs are subtly manipulated to exploit system vulnerabilities. We identify two key properties for an effective auditing framework:

- *Property 1 (Efficacy)*: The framework should stress the system's ability to maintain consistent and reliable decisions, even when confronted with adversarial input manipulations.
- *Property 2 (Stealth)*: The framework should minimally alter the data, creating subtle yet effective attacks that remain credible to human operators.

Methodology. We adapt the adversarial game approach proposed by (Liguori et al. 2021) to generate realistic adversarial examples. In this method, a data generator is trained to create minimally altered versions of real data instances that can effectively deceive the AI system without the changes being apparent to human supervisors. This adversarial game involves two key components: the data generator and the classifier. The data generator produces modified inputs that stay close to the original data but still cause the AI system to misclassify or make erroneous decisions. Simultaneously,

SA	Constraint	DPD	EOD	Bias Measure	Bias Reduction
Age	Unmitigated	12.38%	27.39%	7.25%	–
	Demographic Parity	1.00%	21.87%	2.94%	**4.31%**
	Equalized Odds	0.38%	29.47%	3.66%	3.59%
	TPR Parity	3.25%	26.13%	4.18%	3.07%
	FPR Parity	12.38%	27.39%	7.25%	0.00%
	Threshold Optimizer	9.75%	17.39%	5.21%	3.43%
Gender	Unmitigated	4.14%	20.00%	3.72%	–
	Demographic Parity	4.93%	23.50%	4.40%	-0.68%
	Equalized Odds	0.57%	5.77%	0.87%	**2.85%**
	TPR Parity	2.36%	19.12%	3.05%	0.67%
	FPR Parity	4.36%	33.73%	5.44%	-1.72%
	Threshold Optimizer	0.29%	26.63%	3.29%	0.43%

Table 1: Results for the German Credit dataset on a Decision-Tree Classifier base model

Model	ROC-AUC Real	ROC-AUC Adversarial	Robustness = $\frac{\text{adv}}{\text{real}}$ ROC-AUC
Logistic Regression	86.10	53.90	62.60
Decision Tree	66.03	51.04	**77.29**
Random Forest	90.56	**60.28**	66.56
Support Vector Machine	**95.54**	52.86	55.32
Multi-Layer Perceptron	89.19	55.72	62.47

Table 2: Performance and Robustness Metrics for Various Credit Scoring Models

the classifier is trained to correctly process these adversarial inputs, thereby enhancing its robustness against such attacks. We evaluate the robustness of AI systems by leveraging the trained generator to produce perturbed versions of input data. By comparing predictions on these perturbed inputs with those on the original data, the system's robustness can be quantified. A significant discrepancy in predictions signals a lack of robustness, with the degree of variation reflecting the system's vulnerability. *Experimental settings.* We conducted an experiment to evaluate the robustness of AI models against adversarial attacks using a real test dataset and an adversarial generator. The AI model outputs either a decision or a probability score based on input data instances. The adversarial generator, trained on the same dataset, minimally perturbs data to create adversarial examples that fool the classifier into altering its decision. The experiment includes two steps: first, assessing the AI model's accuracy on the real test dataset; second, generating a perturbed dataset with slight modifications designed to induce different decisions. A robust model keeps consistent performance across both datasets, while a significant drop on the adversarial dataset reveals vulnerabilities. We trained several machine learning models on this dataset and used the same training data to learn the adversarial generator. The results of these robustness tests are summarized in Table 2.

Observations. First, while the Support Vector Machine is the best performer on the real test set with a ROC-AUC of 95.54, it experiences a significant drop in performance on the adversarial test set, with a ROC-AUC of 52.86. This makes SVM not only the second-least effective method on the adversarial set but also the least robust overall, with a robustness ratio of only 55.32. Second, the Random Forest and Multi-Layer Perceptron (MLP) models show comparable performance on the real dataset, with ROC-AUC scores of 90.56 and 89.19, respectively. However, the robustness of the Random Forest model is clearly higher, as indicated by its robustness ratio of 66.56 compared to MLP's 62.47. This suggests that while both models are strong performers on the real dataset, the Random Forest is better able to maintain its performance when faced with adversarial inputs.

Limitations. Methods with inherently low performance on the real test set can achieve a high robustness ratio simply because their performance does not drop significantly in adversarial settings. However, this high robustness score should not be problematic, as methods with low performance would likely be discarded early in the model selection process, rendering their robustness less relevant in practical applications.

Audit Transparency

Authors in (Hupont Torres et al. 2023) categorised the main elements of information of the AI Act regarding technical documentation and divided them into two areas: datasets and AI systems. The authors also identified five state-of-the-art reporting methodologies published between 2018 and 2022 and analysed which elements were covered by each tool. Moreover, they also indicate four degrees of coverage (low/medium/high/missing) regarding the requirements specified in the AI Act. In this project we have further analysed the content of each reporting methodology, as well as the articles regarding technical documentation in the AI Act, in order to generate a reduced set of questions, aiming to create a documentation guiding system for dataset and AI system reporting. Our workflow includes the following steps (i) List all the questions from the five methodologies and map them into the information element categories identified in (Hupont Torres et al. 2023); (ii) Identify those questions classified as high coverage and discard the rest; (iii) Group similar questions. Select a maximum of five questions per information element; and (iv) Add new elements if needed. Table 3 presents the final set of questions for the datasets' documentation guiding system. Taking into account the requirements of the AI Act we decided to add two new information elements to the datasets table to fully address articles

9 and 10, being (i) License and (ii) Restrictions and Risks. Table 4 presents the final set of questions for AI systems reporting. In this case, we adapted some categories to the ones found in the methodologies of reference. In order to simplify the integration of the transparency guiding tool within the *Multidimensional audit tool*, we discarded the option to add a justification text to the questions. Hence, we formulated our questions as a checklist that would guide dataset creators and AI developers to identify the relevant elements to be reported and align with ethical and legal requirements. We decided to move from a binary-kind of answer to a three-level one, giving an intermediate option (for instance, in the case that data and source code are meant to be open in a future). This checklist is intended to be reviewed along the AI-based system life cycle and hence, answers can evolve during this process. As a result, each checklist produces a radar plot depicting the level of transparency fulfillment for datasets and AI systems (see Figure 1b). This visualisation is inspired by the one proposed in (Vinuesa et al. 2020).

Trustworthiness Assessment

The AGQC was intended to be a streamlined version of the EU ALTAI (*Assessment List for Trustworthy AI*), (Independent High-Level Expert Group on Artificial Intelligence 2019). The ALTAI was created by the Independent *High-Level Expert Group On Artificial Intelligence*, established by the European Commission. The AGQC collapsed multiple categories into a single one and separated one category into multiple where deemed necessary. For example, *Human Agency & Oversight* in the ALTAI was stratified into one category titled "Human Oversight" and another titled "Human Autonomy". The AGQC is also 36 questions as opposed to the ALTAI's 63 questions. The major difference is that the AGQC is interactive. It allows users to quickly answer the questionnaire to have a general understanding of their project's AI safety readiness. The ALTAI is a PDF, which requires more user effort to self-score and visualize results. The ALTAI scoring process works by tallying the scores for each category and then normalizing it to account for categories that have more questions than others. Each question is scored out of two, with a strong affirmative being scored a "two", a weak affirmative a "one", and a negative or a neutral a "zero." For example, the question *Does the system have proper mechanisms to allow for accountability with the system through investigations by both internal and third-party committees and allows for redressing?* has answer choices:

- We have established mechanisms that facilitate the AI system's auditability, risk identification, risk management, and transparency by both internal and third-party committees. We also have the ability to redress mechanisms;
- We only have established mechanisms that facilitate the AI system's auditability by internal and/or third-party committees;
- We do not have established mechanisms and the ability to redress certain issues;
- We consider that these issues are not applicable to our case.

The first choice is graded as a "two," the second a "one," and the last two as a "zero." The visualization uses a spider chart, which was chosen for a clear view of the category scores. Importantly, the visualization supports comparing previous answers to the questionnaire to showcase temporal progress. The visualisation is useful for rapidly assessing progress as the compact form and overlaying of previous reports allow for easy readability. All that is required for the temporal progress chart is a previous CSV file of responses to the report, which can run in just one previous report or multiple. To comply with privacy standards, all results are hosted locally in user CSV files and are not hosted on the creator's end.

Multidimensional Audit Tool

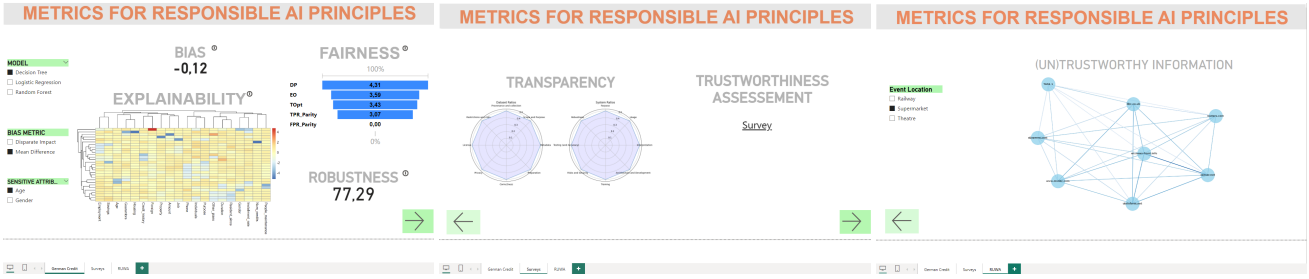
This section presents the interface of the tool, highlighting its main features and visualisations, facilitating interactive data analysis and extracting relevant insights. The *tool* was developed using the *Power BI* platform. The *tool* consists of three pages. The first page, Figure 1a, presents the case study of the German credit data set. The transparency case along with the trustworthiness assessment, are both presented on the second page, Figure 1b, while the case of the RUWA dataset is presented on the third page, Figure 1c.

Discussion

The evaluation and audit aim to measure and assess trust levels through performance analysis indicators, ethical principles or requirements such as transparency and fairness, verification and analysis of algorithms and evaluation metrics to ensure the absence of bias and discrimination, data quality, and compliance with the law. Throughout the article, we present a set of methods and metrics that enable auditing ethics principles under different dimensions: data, model, outputs, predictions, and legal compliance of AI systems for decision-making in two case studies. This tool audits two case studies: the German Credit data set, which focusses on multidimensional aspects such as fairness, explainability, bias, and robustness, and the RUWA data set, which addresses informational bias and transparency in the context of the Russia-Ukraine war. The RUWA case demonstrates that semantic similarity metrics help differentiate between biased news and misinformation. In the German Credit case, bias was confirmed in the data for the sensitive attribute *Age* using the Decision Tree. We observed that the application of constraints led to a considerable reduction in bias, resulting in more fair outcomes, except for *FPRParity*. Our audit presented the best robustness values for the Decision Tree model. Regarding explainability, the attributes *Age*, *Property*, *Num_credits*, *Amount*, and *Job* were identified as determining factors for credit rejection, while *Gender* and *Foreign* had little influence. In terms of evaluation, the results showed that a robust model can present challenges to other technical dimensions, as seen when applying the *ThresholdOptimizer* constraint to decision trees (DT), and that interpretable models do not always produce explainable decisions, requiring post-hoc methods and a participatory design approach to improve, on this case, decision tree explanations. Furthermore, evaluating transparency and other eth-

Information Element	Question
Provenance and Collection	Is the origin of the data public?
	Have you made public the complete data collection process?
Scope and Purpose	Is the scope of the dataset clearly stated?
	Can users know the main purpose for which the dataset was created?
Metadata	Is there an available metadata file dataset comprehension?
Preparation	Is the data processing applied documented and available for users (from the origin to the final dataset)
	Is there a recommended data processing provided?
Correctedness	Have the appropriate tests regarding data completeness been made and published? In the case of incompleteness, is it justified?
	Have you stated if the dataset contains the full population it wants to represent?
	If it does not, have you explained the sampling methods or possible biases?
Privacy	If necessary, have you been transparent regarding the anonymization methods?
License	Is there a document regarding the license of use and distribution of the dataset?
Restrictions and Risks	Are the possible risks of the usage of the data clearly stated in its documentation?

Table 3: Relevant information elements according to AI Act regarding datasets and our documentation guiding system for dataset reporting.



(a) Screenshot of 1st *Dashboard* page (b) Screenshot of the 2nd *Dashboard* page (c) Screenshot of the 3rd *Dashboard* page

Figure 1: Dashboard - Multidimensional Audit Tool

ical components assessed through a questionnaire presents additional challenges when measuring their outcomes. Part of this complexity arises from the proximity to a human-centred approach, which involves humans in the process and evaluation, which is essential for both technical and ethical dimensions, highlighting the need to deepen technical knowledge and promote a robust integration with ethical principles. The experiments carried out show that interventions aimed at mitigating bias do not always preserve the robustness of the system. In high-risk systems, in addition to reporting individual metrics, the effects of mitigation strategies should be documented and the choices justified in light of the AI Act requirements, whenever possible.

Conclusion

This work contributes a novel, integrative framework for ethical AI auditing that (i) aligns with current regulatory demands (e.g., EU AI Act), (ii) combines multiple dimensions (fairness, robustness, explainability, etc.) into a cohesive tool, and (iii) demonstrates practical applicability through two, high-risk case studies. Unlike prior approaches that focus on isolated aspects, our framework enables a holistic view of ethical compliance and its trade-offs, supported by a dashboard.

Metrics and auditing methods are key instruments to operationalise and monitor compliance with the high-level legal requirements established by the AI Act (Gori 2024b). Yet, the AI Act does not contain a legal definition of "metrics", nor does it establish the use of a specific metric or auditing methodologies as legally mandatory. Providers of high-risk AI systems must, therefore, identify how to operationalise and evaluate compliance with legal requirements and duly document, justify and communicate these choices. Pursuant to Annex IV(2)(g) of the AI Act, the technical documentation of High-risk AI systems must include a detailed description of the metrics used to measure both compliance with the requirements set out in Chapter III, Section 2, and the potentially discriminatory impact of their systems. Providers must also justify why the metrics they choose are appropriate for their specific AI systems (Annex IV(4)), taking into consideration the intended purpose of their systems. Choosing metrics that are appropriate for the intended purpose of an AI system is clearly indispensable to ensure that the performance of the system is adequately tested and that, accordingly, potential risks are assessed and mitigated (art. 9(8) and art. 17, AI Act). The documentation and justification of providers' choices concerning metrics is also key to enabling deployers to understand and correctly use AI systems. This is connected to the requirements of transparency, interpretability, and oversight set out in articles 13 and 14 and Annex IV of the AI Act. In this sense, art. 13 of the AI Act requires providers to include in the instructions for use of their High-risk AI systems information on "the level of accuracy, including its metrics, robustness

Information Element	Question
Purpose	Is the main intended purpose and usage of the system available for the user?
	Have you assessed possible misuses of the system and warned potential users about them?
	Are there clear and measurable goals for the system's performance and outcomes?
Usage	Is there clear documentation regarding the protocols and pipelines needed to follow in order to use the system?
	Are the inputs and outputs of the system clearly stated in its documentation?
	Are there examples and use cases provided to illustrate proper usage of the system?
	Is there a support system or contact for users to report issues or seek help?
Interpretation	Have you included the appropriate information to enable the user to interpret the system's output and its implications?
	Does the system include a tool or similar feature to help the user understand the reasoning behind the system's decisions?
	Are there guidelines on how to use the output in decision-making processes?
Architecture and Development	Is the system's architecture documented and accessible to users?
	Have you publicly assessed scalability and flexibility issues and measures?
	Have you published the baselines and pre-trained models used for your system?
Training	Is the training data used for the system clearly documented and available for review?
	Are users provided with information on how the system was trained and its training data limitations?
	Have you stated the hardware and resources needed for the system training?
Risk and Security	Have potential risks associated with the system's use been identified and documented?
	Is there an open process for regularly updating and patching security vulnerabilities?
	Have you documented the measures taken regarding the system's cybersecurity?
Testing and Accuracy	Are the testing protocols and results documented and available for review?
	If there is an ongoing process for monitoring the system's accuracy and performance, is this process open to the user?
	Are there public mechanisms in place to handle errors and inaccuracies when they occur?
Robustness	Has the system been tested for robustness in real-world scenarios? Are those tests public?
	Have you identified, theorized, and published other possible scenarios in which the system might produce incorrect outputs?

Table 4: Relevant information elements according to AI Act regarding datasets and our documentation guiding system for AI systems reporting.

and cybersecurity against which the High-risk AI systems has been tested and validated and which can be expected" and "the performance of the High-risk AI systems regarding specific persons or groups of persons on which the system is intended to be used" (art. 13(3)(ii) and (v), AI Act). Information about the metrics used by providers is particularly relevant for deployers to duly monitor the operation of the HRAIS and to detect and address anomalies, dysfunctions, and unexpected performance (art. 14, AI Act). Overall, the transparency, human oversight and interpretability requirements established by the AI Act demand that providers to put in place the technical and organisational measures necessary to ensure that operational knowledge of High-risk AI systems and their output is available. This means a form of knowledge that allows providers and deployers to monitor whether AI systems comply with the law and, if necessary, intervene on the system, decide not to use the system, or to overcome its output. In this sense, AI metrics can provide both a first-order and a second-order contribution to providers' and deployers' decision-making: first-order, by giving providers and deployers information on the operations of AI systems; second-order, by offering tools to assess the extent to which such information effectively contributes to the interpretability of the system. The cases studies illustrate that the choices as to which metrics and methodologies are the most adequate to support a legally compliant design, development and deployment of AI systems constitute normative choices that are not neutral and, as it were, automatic. In this respect, our research findings confirm the importance of adopting multiple metrics (Thomas and Uminsky 2020) and (AI HLEG 2019). Using a plurality of metrics can contribute to evaluating AI systems from different perspectives and thereby mitigate the trade-offs inherent to specific forms of quantification and measurement of legally relevant aspects of AI systems design, development and deployment. Our research can help appreciate that providers might find themselves having to reckon with the fact that using different metrics leads to results that are, even partially, incompatible. This warms off providers not to passively lean on the results of the metrics they adopt but to use metrics and auditing methodologies critically, situating them in the context of a broader set of considerations. The *Multi-dimensional audit tool* illustrates in a hands-on perspective how the choices as to metrics and auditing methodologies have inherent trade-offs, thereby offering a concrete example of an AI literacy tool that can foster understanding of the complex interplay between normative and technical requirements that providers and deployers must navigate to ensure compliance with the AI Act.

Therefore, in future work, we intend to expand the framework's capacity to evaluate principles and requirements in different high-risk contexts, datasets, and models while maintaining legal compliance within the context of the European Union. Thus, the evaluation and audit presented reinforce the importance of ethical principles and robust methodologies in building reliable systems aligned with the practical and legal needs of the field in question. The work developed for this tool marks the beginning of a resource that can integrate tools for AI auditing systems. This tool might provide an ethical barometer that developers, policymakers, and legal experts can use in the future.

Acknowledgments

The research reported in this work was partially supported by the European Commission-funded project "Humane AI: Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us" (grant #820437) and was realised within the scope of the project EnSafe, with reference 2024.07677.IACDC, co-funded by FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology), as intermediary beneficiary. (https://doi.org/10.54499/2024.07677.IACDC). MM acknowledges the partial support by the project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union – NextGenerationEU. AC acknowledges the partial support by AI4Europe with grant agreement ID 101070000.

Non-Claim Disclosure

The authors are affiliated with various organizations; however, the views expressed in this publication are solely those of the authors and do not represent the views of their respective employers.

References

AI Act. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Http://data.europa.eu/eli/reg/2024/1689/oj.

AI HLEG. 2019. High-level expert group on artificial intelligence. *Ethics guidelines for trustworthy AI*, 6.

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. Fairness and Machine Learning: Limitations and Opportunities.

Bellamy, R. K. E.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; Nagar, S.; Ramamurthy, K. N.; Richards, J.; Saha, D.; Sattigeri, P.; Singh, M.; Varshney, K. R.; and Zhang, Y. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias.

Bellamy, R. K. E.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilović, A.; Nagar, S.; Ramamurthy, K. N.; Richards, J.; Saha, D.; Sattigeri, P.; Singh, M.; Varshney, K. R.; and Zhang, Y. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5): 4:1–4:15.

Ben-Tal, A.; and Nemirovski, A. 1999. Robust solutions of uncertain linear programs. *Operations research letters*, 25(1): 1–13.

Bertsimas, D.; Pawlowski, C.; and Zhuo, Y. D. 2018. From predictive methods to missing data imputation: an optimization approach. *Journal of Machine Learning Research*, 18(196): 1–39.

Bird, S.; Dudík, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.; and Walker, K. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft.

Bishop, C. M. 1995. Training with noise is equivalent to Tikhonov regularization. *Neural computation*, 7(1): 108–116.

Bobek, S.; Korycińska, P.; Krakowska, M.; Mozolewski, M.; Rak, D.; Zych, M.; Wójcik, M.; and Nalepa, G. J. 2024. User-centric evaluation of explainability of AI with and for humans: a comprehensive empirical study. arXiv:2410.15952.

Cao, J.; and Banaji, M. R. 2016. The base rate principle and the fairness principle in social judgment. *Proceedings of the National Academy of Sciences*, 113(27): 7475–7480.

Caton, S.; and Haas, C. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.*, 56(7).

Chen, C.; Seff, A.; Kornhauser, A.; and Xiao, J. 2015. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, 2722–2730.

Cortes, U.; Cortes, A.; and Barrue, C. 2019. Trustworthy AI. The AI4EU approach. In *Artificial Intelligence for Science, Industry and Society, AISIS2019: 21-25 October, 2019, Universidad Nacional Autónoma de México, Mexico City, México*. PoS - Proceedings of Science.

Dablain, D.; Krawczyk, B.; and Chawla, N. 2024. Towards a holistic view of bias in machine learning: bridging algorithmic fairness and imbalanced learning. *Discover Data*, 2(1): 4.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Finlayson, S. G.; Bowers, J. D.; Ito, J.; Zittrain, J. L.; Beam, A. L.; and Kohane, I. S. 2019. Adversarial attacks on medical machine learning. *Science*, 363(6433): 1287–1289.

Foulds, J. R.; Islam, R.; Keya, K. N.; and Pan, S. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 1918–1921. IEEE.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; III, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Commun. ACM*, 64(12): 86–92.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Gori, G. 2024a. *Legal and Computer Rules: An Overview Inspired by Wittgenstein's Remarks*, volume II of *Anthem Studies in Wittgenstein*, 165–182. Anthem Press. ISBN 9781839991394.

Gori, G. 2024b. Legal Protection by Design in the AI Value Chain. What role for AI Metrics? Report for the HumanE AI Net project. Accessed on 19 Aug 2024.

Government of Canada. 2020. Algorithmic Impact Assessment. Technical report, Government of Canada. Accessed: 2024-08-26.

Guidotti, R. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5).

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR)*, 51(5): 1–42.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hofmann, H. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

Huber, P. J. 1965. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 1753–1758.

Hupont Torres, I.; Micheli, M.; Delipetrev, B.; Gómez, E.; and Garrido, J. 2023. Documenting High-Risk AI: A European Regulatory Perspective. *Computer*, 56: 18–27.

Ibrohim, M. O.; Bosco, C.; and Basile, V. 2023. Sentiment analysis for the natural environment: A systematic review. *ACM Computing Surveys*, 56(4): 1–37.

Independent High-Level Expert Group on Artificial Intelligence. 2019. Trustworthy AI Assessment List. Technical report, European Commission.

Kamiran, F.; and Calders, T. 2009. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, 1–6. IEEE.

Kassambara, A.; and Mundt, F. 2020. *Extract and Visualize the Results of Multivariate Data Analyses [R package factoextra version 1.0.7]*.

Khairova, N.; Galassi, A.; Scudo, F. L.; Ivasiuk, B.; and Redozub, I. 2024. Unsupervised approach for misinformation detection in Russia-Ukraine war news. In *Computational Linguistics Workshop at 8th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2024)*, 21–36. CEUR-WS.

Le Scao, T.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A. S.; Yvon, F.; Gallé, M.; et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*.

Leslie, D. 2020. Understanding bias in facial recognition technologies. *arXiv preprint arXiv:2010.07023*.

Liguori, A.; Manco, G.; Pisani, F. S.; and Ritacco, E. 2021. Adversarial regularized reconstruction for anomaly detection and generation. In *2021 IEEE International Conference on Data Mining (ICDM)*, 1204–1209. IEEE.

Liu, Y.; He, H.; Han, T.; Zhang, X.; Liu, M.; Tian, J.; Zhang, Y.; Wang, J.; Gao, X.; Zhong, T.; et al. 2024. Understanding llms: A comprehensive overview from training to inference. *arXiv preprint arXiv:2401.02038*.

Liyanage, U. P.; and Ranaweera, N. D. 2023. Ethical considerations and potential risks in the deployment of large language models in diverse societal contexts. *Journal of Computational Social Dynamics*, 8(11): 15–25.

Microsoft Corporation. 2021. Responsible AI Dashboard. Technical report, Microsoft Corporation. Accessed: 2024-08-26.

Mikolov, T.; Grave, E.; Bojanowski, P.; Puhrsch, C.; and Joulin, A. 2017. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*.

Mishra, S.; Shukla, P.; and Agarwal, R. 2022. Analyzing machine learning enabled fake news detection techniques for diversified datasets. *Wireless Communications and Mobile Computing*, 2022(1): 1575365.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 220–229. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Mucha, H.; Robert, S.; Breitschwerdt, R.; and Fellmann, M. 2020. Towards Participatory Design Spaces for Explainable AI Interfaces in Expert Domains (Short Paper). *First International Workshop on Explainable and Interpretable Machine Learning, XI-ML 2020; 43rd German Conference on Artificial Intelligence (KI 2020)*.

Naiseh, M.; Simkute, A.; Zieni, B.; Jiang, N.; and Ali, R. 2024. C-XAI: A conceptual framework for designing XAI tools that support trust calibration. *Journal of Responsible Technology*, 17: 100076.

Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdl, W.; Vidal, M.-E.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3): e1356.

of the European Parliament, R. E. .; and of the Council of 19 October 2022. 2022. Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

Pagano, T. P.; Loureiro, R. B.; Lisboa, F. V.; Peixoto, R. M.; Guimarães, G. A.; Cruz, G. O.; Araujo, M. M.; Santos, L. L.; Cruz, M. A.; Oliveira, E. L.; et al. 2023. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7(1): 15.

Palumbo, G.; Carneiro, D.; and Alves, V. 2024. Objective metrics for ethical AI: a systematic literature review. *International Journal of Data Science and Analytics*, 1–21.

Panigutti, C.; Beretta, A.; Fadda, D.; Giannotti, F.; Pedreschi, D.; Perotti, A.; and Rinzivillo, S. 2023. Co-design of Human-centered, Explainable AI for Clinical Decision Support. *ACM Transactions Interact. Intelligent Systems*, 13(4).

RAIN Group. 2021. RAI Assessment: Responsible AI. Accessed: 2024-08-26.

Rangapur, A.; Wang, H.; and Shu, K. 2023. Finfact: A benchmark dataset for multimodal financial fact checking and explanation generation. *arXiv preprint arXiv:2309.08793*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Saleiro, P.; Kuester, B.; Stevens, A.; Anisfeld, A.; Hinkson, L.; London, J.; and Ghani, R. 2018. Aequitas: A Bias and Fairness Audit Toolkit. *ArXiv*, abs/1811.05577.

Samek, W.; Wiegand, T.; and Müller, K.-R. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models.

Schuler, D. C.; and Namioka, A. 1993. Participatory Design: Principles and Practices.

Schwartz, R.; Vassilev, A.; Greene, K.; Perine, L.; Burt, A.; and Hall, P. 2022. Towards a standard for identifying and managing bias in artificial intelligence.

Shah, M.; and Sureja, N. 2024. A Comprehensive Review of Bias in Deep Learning Models: Methods, Impacts, and Future Directions. *Archives of Computational Methods in Engineering*, 1–13.

Sinha, A.; Namkoong, H.; and Duchi, J. 2018. Certifying Some Distributional Robustness with Principled Adversarial Training. In *International Conference on Learning Representations*.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11): 8135–8153.

Spinuzzi, C. 2005. The Methodology of Participatory Design. *Technical Communication*, 52: 163–174.

Sundararajan, M.; and Najmi, A. 2020. The many Shapley values for model explanation.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Team, R. C. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Thissen, D.; Steinberg, L.; and Gerrard, M. 1986. Beyond group-mean differences: The concept of item bias. *Psychological bulletin*, 99(1): 118.

Thomas, R. L.; and Uminsky, D. T. 2020. Reliance on metrics is a fundamental challenge for AI. *Patterns*, 3.

UNESCO. 2022. Recomendação sobre a Ética da Inteligência Artificial.

Vinuesa, R.; Azizpour, H.; Leite, I.; Balaam, M.; Dignum, V.; Domisch, S.; Felländer, A.; Langhans, S. D.; Tegmark, M.; and Fuso Nerini, F. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature communications*, 11(1): 1–10.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *International conference on machine learning*, 325–333. PMLR.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding deep learning requires rethinking generalization. ICLR 2017. *arXiv preprint arXiv:1611.03530*.