



# Objective metrics for ethical AI: a systematic literature review

Guilherme Palumbo<sup>1</sup> · Davide Carneiro<sup>2,3</sup> · Victor Alves<sup>4</sup>

Received: 19 January 2024 / Accepted: 20 March 2024 / Published online: 13 April 2024  
© The Author(s) 2024

## Abstract

The field of AI Ethics has recently gained considerable attention, yet much of the existing academic research lacks practical and objective contributions for the development of ethical AI systems. This systematic literature review aims to identify and map objective metrics documented in literature between January 2018 and June 2023, specifically focusing on the ethical principles outlined in the Ethics Guidelines for Trustworthy AI. The review was based on 66 articles retrieved from the Scopus and World of Science databases. The articles were categorized based on their alignment with seven ethical principles: Human Agency and Oversight, Technical Robustness and Safety, Privacy and Data Governance, Transparency, Diversity, Non-Discrimination and Fairness, Societal and Environmental Well-being, and Accountability. Of the identified articles, only a minority presented objective metrics to assess AI ethics, with the majority being purely theoretical works. Moreover, existing metrics are primarily concentrating on Diversity, Non-Discrimination and Fairness, with a clear under-representation of the remaining principles. This lack of practical contributions makes it difficult for Data Scientists to devise systems that can be deemed Ethical, or to monitor the alignment of existing systems with current guidelines and legislation. With this work, we lay out the current panorama concerning objective metrics to quantify AI Ethics in Data Science and highlight the areas in which future developments are needed to align Data Science projects with the human values widely posited in the literature.

**Keywords** Artificial Intelligence · Ethics · Metrics · Systematic review

## 1 Introduction

Artificial Intelligence (AI) is widely recognized as a significant disruptive force across all domains it touches. The widespread adoption of AI has experienced significant growth, resulting in a substantial impact on society, potentially with both positive and negative consequences. While AI

is expected to be immensely beneficial for humanity, namely in areas such as medicine, law, education, or industry, its negative impacts may outweigh the positive ones if it is not developed in a way that has human values at its core.

Early examples of the potential risks abound, such as AI systematically discriminating against black patients by miscategorizing them in heart failure risk scores or kidney donor risk indexes, marking black individuals as less suitable donors [1], or AI recruiting tools that show bias against women [2] or even a chatbot coaching a “girl of 13 years old” on losing virginity [3].

In order to avoid such cases of harmful AI, legal and ethical regulation is paramount. However, one challenge when it comes to regulation is that legislation will always move slower than technological development. While the European Commission’s risk-based approach on an AI legislation is a step in the right direction, in the sense that it is not based on specific technologies or systems but on their level of potential risk, it might not be enough.

We argue that in order to allow AI developers and Data Scientists to actually comply with legislation, transparency requirements, and to be aware of the potential issues in the

Guilherme Palumbo, Davide Carneiro and Victor Alves have contributed equally to this work

✉ Guilherme Palumbo  
gfp@estg.ipp.pt

✉ Davide Carneiro  
dcarneiro@estg.ipp.pt

Victor Alves  
valves@di.uminho.pt

<sup>1</sup> CIICESI, ESTG, Politécnico do Porto, Felgueiras, Portugal

<sup>2</sup> INESC TEC, Rua Dr. Roberto Frias, 712, 4200-465 Porto, Portugal

<sup>3</sup> ESTG, Politécnico do Porto, Felgueiras, Portugal

<sup>4</sup> ALGORITMI Research Centre / LASI, University of Minho, Braga, Portugal

applications they develop, there is a need for more practical tools [4, 5]. Tools that go beyond checklists or toolkits that can pinpoint sources of potentially ethical issues at any stage of the Data Science lifecycle, preferably before they have an actual impact. Without such an objective approach, we regard it as very difficult if not impossible to argue transparently about the level of ethical compliance or alignment of a given system or to point out ethical issues so that they can be promptly addressed.

One first challenge thus concerns the choice of ethical principles to be considered in order to create AI systems that are ethically aligned and cause no harm to society. This has been addressed by the High-Level Expert Group on AI, which in 2019 defined the Ethics Guidelines for Trustworthy Artificial Intelligence [6] by presenting several principles an AI must adhere to in order to be deemed trustworthy and ethical. These guidelines were developed in order to make AI lawful (respecting all applicable laws and regulations), ethical (respecting ethical principles and values) and robust (both from a technical perspective, while taking into account its social environment).

In this paper, when referring to ethical principles, we adhere to the work of the Expert Group on AI. According to its Ethics Guidelines, there are a set of seven key requirements that AI systems should meet in order to be deemed trustworthy:

- **Human agency and oversight:** AI should empower humans to make educated decisions and uphold their rights. While ensuring sufficient oversight, human-in-the-loop, human-on-the-loop and human-in-command methodologies can be used.
- **Technical Robustness and safety:** AI must be secure and resilient. It must be accurate, dependable, reproducible and safe, with a backup plan. That is the only method to reduce and prevent unintentional injury.
- **Privacy and data governance:** In addition to privacy and data protection, data governance procedures must consider data quality and integrity and legitimize data access.
- **Transparency:** Data, systems and AI business models should be transparent. Traceability methods aid in this. Additionally, stakeholders should be informed about AI systems and their conclusions, since AI needs to be explainable. People must know when they are interacting with an AI system and its capabilities and limitations.
- **Diversity, non-discrimination and fairness:** AI must avoid unfair bias, which can marginalize vulnerable populations and worsen prejudice and discrimination. AI systems should be accessible to all, regardless of disability, and involve stakeholders throughout their lives to promote diversity.

- **Societal and environmental well-being:** AI should help all humans, including future generations. It must ensure they are sustainable and eco-friendly. They should also consider the environment, including other living things, and their social and societal influence.
- **Accountability:** Responsibility and accountability for AI systems and their results should be established. Auditability, which allows algorithm, data and design process evaluation, is crucial in important applications. Additionally, a proper and accessible remedy is needed.

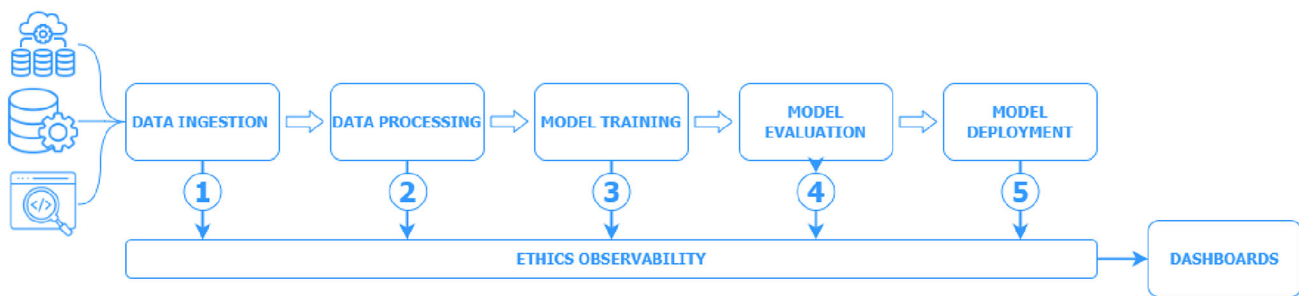
A second challenge is, once the relevant principles are established, how to measure an AI system's level of compliance or alignment with these principles. In our view, this must go beyond mere words or statements of commitment by organizations, but include automated, objective and standard assessment means, and to communicate the results of these internal analysis to the public. Such an approach should guide the development of AI in light of generally accepted human-centric values and principles.

However, principles by themselves are very high-level and abstract by nature. They are, *per se*, of low usefulness to Data Scientists or Computer Scientists, who work at a technical level, and may struggle to find effective ways to implement them. Moreover, these principles are also very much subject to interpretation, and different organizations or individuals may interpret them differently and implement them differently, eventually leading to a lack of standardization and unfair competition. Ultimately, this may generate confusion and a barrier to information for the end-user.

In this paper, we argue that abstract and high-level concepts and guidelines, while being an essential starting point, are not enough. We believe that only with the creation of objective and observable ethics metrics that can be integrated into a generic Data Science pipeline, will we be able to: 1) have a real-time perception of the level of alignment of a given system; 2) pinpoint its root causes (e.g., data, model, processes); and 3) take the appropriate counter-measures to, first, avoid a negative impact on the users and, secondly, address the root cause.

Especially important is the notion that this is not a one-shot intervention: a system that is deemed ethically aligned in a given moment in time may not be deemed so at a later stage. This can happen, for instance, due to changes in the parameters or hyper-parameters of a Machine Learning (ML) model (that drove it to overfit some particular data or become biased), or to changes in the underlying data (that became itself biased or of poor quality). So, monitoring ethical compliance through objective metrics or indicators must be seen as a continuous effort throughout all the stages of any Data Science pipeline.

Our contribution to address this challenge lies in what we consider the first necessary step: to carry out a survey and



**Fig. 1** Artificial Intelligence data pipeline with ethics observability solution

characterization of existing objective metrics that can be used to quantify, to some extent or in some dimension, the level of ethical alignment of a Data Science pipeline. To do so, this paper presents the results of a systematic literature review on metrics-based ethical assessment and compliance in AI systems. The goal is to carry out a survey of existing objective ethics metrics and to organize them by ethical principle in order to create a map of how each principle is currently measurable and observable in a typical Data Science pipeline, according to the literature.

In the context of this paper, when using the term *metrics*, we refer to any measure in which an objective quantitative or qualitative value is assessed in relation to a standard or scale. These metrics yield specific numerical values, percentages, frequencies, or even functions that measure the distance between two outcomes within a defined space while also being directly linked to the evaluation of the ethical principles of AI. A metric can also be a qualitative measure that shows non-quantitative conformance to a specific criterion, also directly related to the evaluation of the ethical principles of AI.

This SLR was designed with the goal to find practical and objective metrics that could be seamlessly integrated in any typical data pipeline, such as the one represented in Fig. 1. This pipeline, depicted here in very general terms, also conveys another important criterion: the metrics identified should be technology-agnostic and re-usable in any data pipeline. They should be relevant and applicable to any set of data, be it streaming or batch. We want the identified metrics to be useful for the development of monitoring systems that will analyze, in real time, the ethical compliance of a data pipeline under the principle of observability. This integration is crucial for practical implementation, enabling data scientists and AI developers to incorporate ethical considerations into their models and algorithms from the onset. By focusing only on quantifiable and automated metrics, the aim is to provide concrete, actionable insights in real time that align with the data-driven nature of AI and DS systems.

For this reason any tool, metric, framework, checklist, toolkit, or theoretical concept that either require human inter-

vention (and may thus be prone to bias or individual choices) or cannot be implemented in a fully automated way were deliberately left out regardless of their value in the ethical AI landscape.

Indeed, tools or frameworks that need human intervention can vary significantly in their application across different contexts and domains, making it challenging to establish a standardized set of metrics. Additionally, relying on manual metrics or metrics that aren't useful for explaining how ethical the AI is introduces subjectivity and inconsistency, which may undermine the reliability of the findings. Furthermore, by eschewing existing assessment frameworks or manual metrics that need human intervention, the risk of bias or preconceived notions influencing the selection of metrics is avoided. Instead, it was adopted a more agnostic approach, focusing solely on the quantifiable metrics of ethical AI that could be implemented and automated into the standardized AI or DS pipeline. While this approach may limit the scope of the study, we believe it provides a solid foundation for future research and practical implementation.

It is also not our goal to identify solutions or mitigation approaches to ethical problems, such as tools that do automatic resampling of data or bias mitigation tools [7]. We do, however, identify metrics or indicators that can identify and quantify such problems.

The rest of the paper is organized as follows: Sect. 2 presents the research goals and methodology of the systematic literature review (SLR); Sect. 3 presents the results of the SLR; Sect. 4 addresses the main findings derived from the SLR and; Sect. 5 presents the conclusion of this SLR.

## 2 Research goals and methodology

The main goal of this SLR is to identify objective ethics metrics currently present in academic literature. Moreover, we also aim to create a map of how the identified metrics cover the relevant ethical principles proposed in the literature in order to determine whether additional developments are necessary and, if so, in which areas.

Systematic reviews are characterized by a methodical and replicable methodology and presentation. They involve a component of comprehensive search to locate all relevant published work on a subject, followed by a systematic integration and synthesizing of the search results and a critique of the extent, nature and quality of evidence in relation to a particular research question, highlighting gaps between what is known and what needs to be researched/developed [8].

This SLR was elaborated following PRISMA Method (Preferred Reporting Items for Systematic Reviews and Meta-Analysis). PRISMA considers a set of evidence-based minimum items for reporting systematic reviews and meta-analysis, especially for evaluating interventions [9]. It primarily focuses on the reporting of reviews evaluating the effects of interventions, but can also be used as a basis for reporting systematic reviews with objectives other than evaluating interventions.

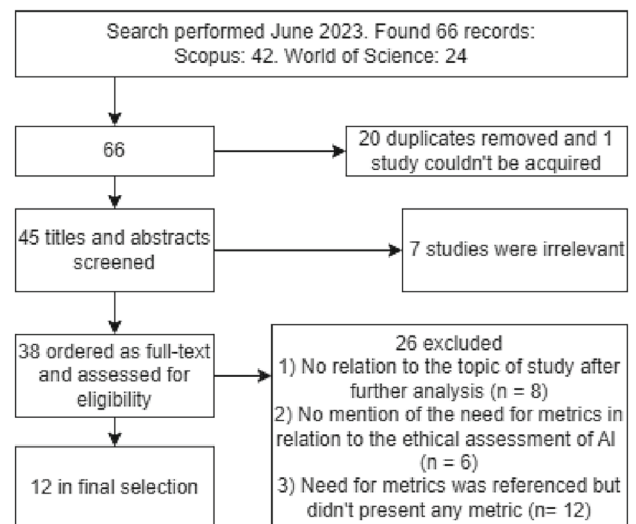
The first stage of this SLR started by identifying the research issues. In this instance, and after a first comprehensive search, any paper related to the assessment of the ethics of AI based on metrics was considered eligible. So, any study, article, journal, or paper whose title, abstract, or keywords included terms related to ethics, Artificial Intelligence and metrics was obtained. In summary, the search keywords were “Ethics” AND “Artificial Intelligence” AND “Metrics,” and the query used to extract any related paper was: (TITLE-ABS-KEY (metrics) AND TITLE-ABS-KEY (artificial AND intelligence) AND TITLE-ABS-KEY (ethics)) AND PUBYEAR > 2017.

This SLR was conducted in June 2023; consequently, any additional papers introduced after that date have not been taken into account.

Hence, the available literature related to the metrics-based assessment of the ethical compliance of AI was reviewed for a comprehensive understanding of which solutions and research have already been conducted.

The flowchart depicted in Fig. 2 summarizes the process followed for the retrieval of the literature. In total, the literature search yielded 66 papers. Of these, 45 papers remained after the removal of duplicates (20) and the impossibility of accessing 1 paper. This specific paper couldn't be accessed due to its cost. However, the analysis of its abstract also revealed that it would not be relevant given the goals and scope of this literature review.

The second stage consisted of identifying relevant studies based on the preceding findings and selecting the studies that were relevant to the subject. Although the relevance of ethics metrics and principles varies according to the domain (e.g., healthcare, law, education), we approach this work from a domain-agnostic perspective, as the implementation of each metric is the same regardless of the domain. For this reason, we did not filter papers based on their domain.



**Fig. 2** Flow chart summarizing the process of paper selection for inclusion in the study

The search was carried out on two major databases, Scopus and World of Science, and this search was intended to acquire all related studies, articles and academic papers in the last five years, starting in January 2018 and finishing in June 2023. This date interval was defined for several reasons. First and foremost, we wanted to identify recent relevant contributions since the explosion in AI research that started around 2018. Second, while we acknowledge that there is relevant literature on AI Ethics before 2018, the existing work is largely theoretical or fundamental and thus not aligned with the goals and scope of this literature review.

Furthermore, we also did not restrict the search to a specific geography. We acknowledge that the European space is, arguably, at the forefront of AI Ethics and regulation, and the work developed is aligned with the principles proposed by the European Commission. However, any research anywhere can make a valuable contribution.

For a similar reason, we did not include in this search documents of non-scientific nature, such as reports generated or requested by political entities, legislation, or guidelines. While relying only on the outputs of scientific research, we believe to have a guarantee of impartiality and for unbiased results.

Duplicated papers were removed at this stage. Based on the topics covered by their titles and abstracts, 38 articles were obtained as full-text and assessed for eligibility [10–47].

The selection criteria were established so as to eliminate those articles that, after an analysis of their abstracts, were in no way related to the topic in question. Therefore, any article that didn't address ethical AI and/or metrics in some way, not even in an implied way, was excluded.



Consequently, following a full-text screening of each remaining paper, every paper that discussed the need for metrics, evaluation, and analysis of any ethical principle, even if only theoretically or not at all explicitly, was selected. This allowed to quantify the frequency with which the evaluation of each principle was addressed in the literature, thus allowing to estimate its level of maturity. Then, an in-depth full-text analysis of the previously selected pool of papers was carried out. Any article that didn't present at least a single practical and objective metric to assess an AI system's level of compliance with any of the stated principles was removed from the analysis.

In the third and final stage, data from the selected papers were collected, systematized and summarized, and the results were reported. Initially, the relevant information was extracted and summarized on a per-article basis. So, a detailed summary was elaborated for each selected paper based on a full-text analysis. Specifically, the focus was put on relevant information regarding the domain of application, ethical principle, solution provided, metric proposed (if any), specific outcomes, among others. In this stage, papers were also grouped according to the ethical principle(s) addressed (e.g., bias, robustness, explainability), in order to ascertain the level of coverage of each principle in the literature.

Out of the 38 papers eligible for full-text assessment, 8 didn't fit any subject of the literature review and were considered irrelevant [35, 36, 38, 40, 44, 46, 47]. The exclusion of these 8 papers was carried out on the grounds that they had no relationship with the topic of study, since they didn't address ethical AI nor address or mention any kind of metrics for the ethical assessment of AI.

Another group of 6 papers [34, 37, 39, 41, 42, 45] in the pool of full-text assessment discussed ethical AI but didn't address or mention the need for the evaluation or the need for metrics for the ethical assessment of AI. For instance, Edwards [34] addresses the IEEE P7010 standard and how it can help support organizations and individuals who develop or use AI in considering and integrating ethical considerations for human well-being in product development. This article (in theory) discusses how IEEE P7010 presents several metrics of growth, income and productivity of well-being, but in practice, this article didn't present any relevant metric that is quantifiable, measurable, or that evaluates how ethical an AI is; the article solely made reference to the fact that they exist without showing any relevant proof.

Similarly, Germann et al. [47] intend to clinically validate a Deep Convolutional Neural Network (DCNN) for the detection of surgically proven Anterior Cruciate Ligament (ACL) tears in a large patient cohort and to analyze the effect of magnetic dissonance examinations from different institutions with varying protocols and field strengths. The paper validated the DCNN and concluded that it showed sensitivities and specificities well above 90% for the entire

population, irrespective of the magnetic field strength of the MRI. In this case, this paper is not relevant since it doesn't address the ethical analysis of the model nor does it present any relevant ethics metric.

Finally, 12 papers [15, 17, 21, 24, 25, 27–33] explored the necessity of metrics and evaluation methods for assessing the ethical implications of AI. However, these studies did not provide specific metrics and/or rather proposed theoretical solutions that relied on checklists or toolkits.

For instance, Saetra et al. [24] present an approach to the evaluation of the positive and negative impacts of AI on sustainable development (regarding ESG (Environment, Social, Governance) reporting to the United Nations Sustainable Development Goals (SDG)). It focuses on the examination of micro-, meso- and macro-level impacts, considering both negative and positive impacts, and accounting for ripple effects and inter-linkages between the different impacts. Although in good spirit and relevant, the article never presented a solution based on metrics nor presented a way to evaluate the ethical state of an AI system, making it out of the scope of this work.

As previously mentioned, the 38 papers were grouped according to the ethical principles mentioned in each one, with the resulting clusters allowing to quantify the frequency of each principle in the literature (independently of the proposal of actual metrics or not). Furthermore, a total of 12 papers [10–14, 16, 18–20, 22, 23, 26] fulfilled all the inclusion criteria, including presenting at least one ethics metric, and were selected for a more in-depth analysis. The results of this analysis are detailed in Sect. 3.

## 3 Results

### 3.1 Literature summarization

Raji and Buolamwini [26] investigate the commercial impact of algorithm audits (Gender Shades) on increasing algorithmic fairness and transparency. The Gender Shades study audits commercial facial recognition products to assess their ability to correctly identify gender and skin types. The auditors (an independent group) follow a procedure similar to coordinated vulnerability disclosures in information security, which involves documenting the vulnerabilities they find and giving the companies a chance to respond before publicly releasing the result. The goal is to expose performance vulnerabilities in commercial facial recognition. The study targets IBM, Microsoft, Megvii, Amazon, Kairos, etc. Based on the measurement of error differences across the identified subgroups, they conclude that all targets reduce accuracy disparities between males and females and darker and lighter-skinned subgroups, with the most significant update occurring for the dark-skinned female subgroup,

which undergoes a 17.7–30.4% reduction in error between audit periods. In this study, a series of ethics metrics are presented.

Kaul and Soofastaei [10] address the current state of AI ethical principles in the mining industry. Later, they present a series of guidelines and recommendations on how to use AI ethically across the project lifecycle (CRISP-DM). The goal is also to help the organization understand, evaluate and mitigate bias using the appropriate fairness metrics. They present some fairness metrics and bias mitigation algorithms to help remove bias (e.g., equal opportunity, demographic parity, disparate impact, Theil index, among others).

Kasirzadeh [11] addresses feminist political theory (the work of Iris Marion Young) and argues that a primarily distribution approach to the local allocation of material and computational goods is too narrow to adequately address or accommodate concerns about social injustice vis-à-vis algorithmic fairness. It argues that algorithmic fairness is morally salient and requires philosophical attention and that algorithmic ecosystems are socio-technical entities and are therefore receptive to different sources of social injustice. It argues that the metrics are just concerned with local matters of distributional justice. However, not all sources of social injustice are distributional, some are structural. The paper also argues for six positive corollaries of the adoption of socially responsible algorithmic fairness as the conceptual basis for research into the infrastructural fairness of algorithmic ecosystems and their direct effects. In summary, the paper attempts to connect some dimensions of the philosophical works of Young to algorithmic fairness. In the process, it discusses counterfactual comparison, demographic parity, equal opportunity, etc.

Zafar et al. [12] focus on an algorithm for diabetic retinopathy screening and risk stratification. According to the study, additional performance metrics are needed that extend beyond the assessment of technical accuracy in order to comprehensively understand the influence of AI algorithms on patient outcomes. There is a need for real-world evaluation of safety, efficacy and equity (bias), impact of patient outcomes, ethical evaluation (using federated learning to test privately against the same algorithm rather than using pooled data), logistical and regulatory evaluation. In summary, the article focuses on arguing that there is a need for real-world validation and testing. It also makes reference to equal opportunity and equalized odds when addressing bias, arguing that there are two types of AI bias that are well defined: the (in)equality of opportunity (equal precision) and the (in)equality of odds (equal false positive and false negative rates) that can occur at all stages of the development of AI algorithms.

Bae and Xu [13] present two state-of-the-art pedestrian trajectory prediction models for age and gender bias across three different datasets (JAAD, PIE, and TITAN). The goal is to design and utilize novel evaluation metrics for compar-

ing model performance (mean MSE, Mann–Whitney U test, Wasserstein distance). Both models (BiTraP, SGNet) perform worse on children and the elderly compared to adults. The paper also identifies potential sources of biases and some metrics to identify them (demographic parity/statistical parity), as well as discussing several limitations of the study. It concludes that there is no clear difference between genders.

Kasirzadeh and Clifford [14] lay the foundation for a more detailed analysis of the fairness principle by first exploring the potential for fairness metrics to operationalize the principle in order to more adequately respond to the potential for unfair outcomes. The paper also argues that there is a significant limitation to the use of fairness metrics to process personal data fairly. It discusses popular metrics for the assessment of algorithmic fairness, such as statistical parity and equality of opportunity, and also one way to provide a more concrete analysis of the notion of individual fairness by using counterfactuals. It concludes that the technical challenges have an impact on the usefulness of Data Protection Impact Assessments irrespective of a controller's willingness to actively engage in the process. It also concludes that the fairness metrics are context dependent and have varying interpretations of fairness according to different fairness metrics. Additionally, it concludes that data controllers play a key role in the determination of what is fair.

Cortés et al. [16] introduce the notion of locality and defines structural interventions. They compare the effect of structural interventions on a system compared to local, structured-preserving interventions on technical objects. The paper proposes a methodology (RISE) to account for elements of algorithmic discrimination based on social origin. It places the algorithm in the social context in which it is deployed instead of just considering the algorithm in isolation. The framework allows for the identification of bias outside the algorithm stage and proposes joint interventions on social dynamics and algorithm design. To evaluate this proposal, they use demographic parity, equal opportunity and equalized odds metrics. The paper concludes by showing several structural interventions in a model for financial lending and concludes that structural interventions, unlike algorithmic interventions, can in fact help the system become more equal.

Zhang et al. [18] address the existing conflicts and inconsistencies among accuracy and fairness metrics. The paper considers how to simultaneously optimize accuracy and multiple fairness metrics more effectively. Not only that, but the paper presents 16 fairness metrics and, based on the obtained correlations, concludes that 8 fairness metrics represent all 16. The paper also views the mitigation of unfairness as a multi-objective learning problem, and a multi-objective evolutionary learning framework is used to optimize the metrics simultaneously. Afterward, ensembles are constructed based on the models in order to automatically balance different

metrics. The authors also analyze the approach in eight different datasets and conclude that the framework can improve fairness according to a broad range of fairness metrics, even the ones not used in the multi-objective learning algorithms. They present the results and conclude the model performs well for both the eight fairness metrics used in training (average odd difference, error difference, discovery ratio, predictive equality, false omission rate difference, false omission rate ratio, false negative rate difference, false negative rate ratio) and for the other eight fairness metrics not used in training (error ratio, discovery difference, false positive rate ratio, disparate impact, statistical parity, equal opportunity, equalized odds, predictive parity).

Schedl et al. [19] address the necessity of discussing with different stakeholders when investigating biases and fairness in the value chain of recommendation systems. It also points out that there are discrepancies between computational metrics of bias and fairness (disparate impact, generalized entropy index, statistical parity) and their actual individual and societal perception. It points out that bias cannot be measured in a fully objective way, but it pushes for a more holistic perspective on human perception in relation to psychological, sociological and cultural backgrounds. It finally discusses metrics of bias and fairness, as well as technical debiasing solutions in the context of ethical considerations and legal regulations.

Goethals et al. [20] address the use of counterfactual explanations to assess the fairness of a model. Firstly, the paper presents some fairness metrics (demographic parity, disparate impact, equal opportunity, equalized odds, statistical parity) and counterfactual fairness. The paper also argues that the counterfactual explanations can not only detect explicit bias (when using a sensitive attribute) but also implicit bias (when not using the sensitive attribute). It presents the Pre-CoF metric (Predictive Counterfactual Fairness) which is successfully used to detect implicit bias in the model in the use cases.

Fleisher [22] argues that the method of individual fairness does not serve as a definition of fairness and does not serve as the only method analyzed, nor should it be given priority over other fairness methods. In the process, the paper addresses equalizing odds, measuring statistical distance and achieving parity. The author presents four in-principles problems for individual fairness, which are the insufficiency of similar treatment (showing similar treatment is insufficient to guarantee fairness), systematic bias and arbiters (the method is at risk of encoding human implicit bias), prior moral judgments (individual fairness requires prior judgment, limiting its usefulness as a guide for fairness), and the incommensurability of relevant moral values makes similarity metrics impossible for many tasks (two values are incommensurable if there is no common measure that can be applied to both values).

Wachter et al. [23] address the critical gap between legal, technical and organizational notions of algorithmic fairness. By analyzing EU non-discrimination law and the jurisprudence of the European Court of Justice (ECJ) and national courts, it identifies a critical incompatibility between European notions of discrimination and existing work on algorithmic and automated fairness. There is a clear gap between statistical measures of fairness (e.g., demographic parity, equalized odds) embedded in various fairness toolkits and governance mechanisms and the context-sensitive, often intuitive, and ambiguous discrimination metrics and evidential requirements used by the ECJ. The article makes three contributions. First, the authors review the evidential requirements to bring a claim under EU non-discrimination law. Due to the disparate nature of algorithmic and human discrimination, they conclude that the EU's current requirements are too contextual, reliant on intuition and open to judicial interpretation to be automated. Secondly, they show how the legal protection offered by non-discrimination law is challenged when AI, not humans, discriminates. Compared to traditional forms of discrimination, automated discrimination is more abstract and unintuitive, subtle, intangible and difficult to detect. Thirdly, they examine how existing work on fairness in machine learning lines up with procedures for assessing cases under EU non-discrimination law. The paper proposes 'conditional demographic disparity' (CDD) as a standard baseline statistical measurement that aligns with the Court's 'gold standard'.

This concludes a summary of each paper's contribution. The following sub-sections describe and define the Ethical principles considered in this work, as well as the identified metrics.

### 3.2 Ethical principles

Over the past years, as a result of the need for more human-aligned AI development, many principles have been put forward and many subjects have become widely discussed. Table 1 systematizes the key subjects identified in the 38 papers analyzed in the context of this SLR, as well as on the Ethics Guidelines for Trustworthy AI [6].

Based on the literature review, the most frequently addressed subject is Bias, which is mentioned in 22 papers (58%), closely followed by Fairness, which is mentioned in 20 studies (53%). Other frequent subjects include Transparency, referred in 12 studies (32%), and Privacy, mentioned in 11 studies (29%).

Precisely defining and distinguishing these subjects is also often challenging, as some of them overlap significantly according to the accepted definitions in the literature. This happens because they have been proposed freely by researchers, using terms that are sometimes synonyms or very closely related. Since having a large number of (potentially

**Table 1** Frequency in which the identified subjects were addressed in the 38 papers, in descending order

Subject	Frequency (%)
Bias	22 (58%)
Fairness	20 (53%)
Transparency	12 (32%)
Privacy	11 (29%)
Accountability	7 (18%)
Governance	6 (16%)
Explainability	6 (16%)
Well-being	4 (11%)
Human Agency	2 (5%)
Non-maleficence	2 (5%)
Safety	2 (5%)
Robustness	1 (3%)

overlapping) key subjects might dilute the findings of the work to be carried out, a decision was taken to cluster every subject within a specific ethical principle based on the Ethics Guidelines for Trustworthy AI [6].

Thus, Human agency was attributed to the ethical principle of *Human Agency and Oversight*; Robustness and Safety were clustered into *Technical Robustness and Safety*; Privacy and Governance were assigned to *Privacy and Data Governance*; Transparency and Explainability to *Transparency*; Bias and Fairness were clustered into *Diversity, Non-Discrimination and Fairness*; Well-being and Non-maleficence were assigned to *Societal and Environmental Well-being*; and Accountability into *Accountability*.

This clustering allows to better understand the attention that has been devoted by the literature to each ethical principle, as detailed in Table 2, which presents the frequency of each ethical principle considering the 38 articles. Table 3, on the other hand, organizes the 38 papers according to the ethical principles they address. The results are comparable to the previous ones, where the principle that was devoted the greatest attention was *Diversity, Non-Discrimination and Fairness*, having been mentioned in 24 different papers (58%), followed by Transparency being addressed in 14 different studies (37%) and Accountability with 7 presences (18%).

Clearly, this analysis shows an imbalance between principles (dominated by Diversity, Non-Discrimination and Fairness). However, according to the goals of this work, we deepen this analysis by examining which papers provide tangible outcomes and practical metrics for assessing the ethical condition of IA, and not merely mention a given principle.

As detailed in Table 4, only 12 articles out of 24 related to Diversity, Non-Discrimination and Fairness presented in fact at least 1 actual practical metric. Moreover, these articles

**Table 2** Frequency in which an Ethical AI Principle was addressed in the 38 articles, in descending order

AI ethical principle	Frequency (%)
Diversity, non-discrimination and fairness	24 (58%)
Transparency	14 (37%)
Accountability	7 (18%)
Privacy and data governance	6 (16%)
Societal and Environmental Well-being	6 (16%)
Technical Robustness and Safety	3 (8%)
Human agency and oversight	2 (5%)

correspond to 100% of all articles that presented metrics and to 34% of all full-text analyzed articles.

Thus, several key findings can already be drawn:

- Only a small amount of the articles (34%) provide an actual objective and measurable metric;
- All the articles that mention objective metrics (12) relate to the principle of “Diversity, Non-Discrimination and Fairness”;
- There is a significant need to investigate and propose novel metrics that can be used to quantify the level of compliance of an AI system with the remaining ethical principles.

Table 5 identifies the 12 articles that propose actual metrics, by principle. The following section discusses in detail each of the identified metrics in these 12 articles.

### 3.3 Ethics metrics

After the analysis of the 12 studies that present actual objective metrics, it is imperative to understand the specific concept measured by each metric and how this is accomplished, to comprehend their relevancy.

Similar to the ethical principles, it was found that many metrics are similar or even overlap in the concept they represent, despite having slightly different names. Thus, a similar exercise was conducted, in which groups of equivalent metrics were clustered into a single one, so as to prevent repetitions and better focus the analysis. Table 6 shows the result of this clustering. The left column shows all the metrics found in the literature. These metrics have then been clustered according to their meaning, as depicted in the right column.

In total, after their clustering, 15 objective metrics have been identified in the articles. Independently of their names in the original papers, they will be, from this point onward, referred to by the name on the right column of Table 6.



**Table 3** Analyzed papers grouped by the addressed ethical principle

AI ethical principle	Articles
Diversity, non-discrimination and fairness	[10–14, 16, 25–30, 41, 42] [18–23, 31, 32, 39]
Transparency	[12, 14, 15, 17, 20, 24, 26, 28, 30, 31, 37, 41, 45] [33]
Accountability	[14, 17, 26, 28, 31, 39, 41]
Privacy and data governance	[15, 24, 26, 28, 33, 39]
Societal and environmental well-being	[24, 26, 28, 31, 33, 34]
Technical robustness and safety	[26, 30, 37]
Human Agency and Oversight	[17, 31]

**Table 4** Frequency in which an Ethical AI Principle was addressed in the final 12 studies, the ones that present at least one metric for the ethical assessment of AI, in descending order

AI ethical principle	General frequency (%)	Frequency based on the final selection (%)
Diversity, non-discrimination and fairness	12 (34%)	12 (100%)
Transparency	0 (0%)	0 (0%)
Accountability	0 (0%)	0 (0%)
Privacy and data governance	0 (0%)	0 (0%)
Societal and environmental well-being	0 (0%)	0 (0%)
Technical robustness and safety	0 (0%)	0 (0%)
Human agency and oversight	0 (0%)	0 (0%)

**Table 5** Related articles that presented at least one metric for the ethical assessment of AI to its corresponding Ethical AI Principle in the final 12 studies, in descending order

AI ethical principle	Articles
Diversity, non-discrimination and fairness	[10–14, 16, 18–20, 22, 23, 26]
Transparency	
Accountability	
Privacy and data governance	
Societal and environmental well-being	
Technical robustness and safety	
Human agency and oversight	

As it's clear to see, the majority of the metrics used to assess the level of ethical compliance of AI are fairness metrics. Fairness in machine learning refers to the various attempts at correcting algorithmic bias in automated decision processes based on machine learning models and identifying historical systematic disadvantages.

According to the literature, there is a lack of consensus in the community about what is considered bias or fair and a lack of consensus among different measures. Since fairness in different contexts can be interpreted into different quantitative definitions to emphasize different perspectives, no single measure has been accepted as a universal notion of fairness quantification [18, 19].

There are several fairness metrics that, depending on the context, are relevant and useful to mitigate and identify

bias. Fairness metrics can be defined into two conflicting but complementary categories: group fairness and individual fairness.

Group fairness is the idea that the average classifier behavior should be the same across groups defined by protected attributes, while individual fairness is the idea that all individuals with the same feature values should receive the same predicted label and that individuals with similar features should receive similar predicted labels. Individual fairness includes the special case of two individuals who are the same in every respect except for the value of one protected attribute [22, 48].

Typically, protected (also called sensitive) attributes are traits considered to be discriminative by law, such as gender, race, age, among others [18].

As detailed in Table 6, a total of 33 distinct metrics were identified; however, not all of these mentioned metrics are unique; rather, some are identical and were merely designated by different names in different papers. In the remainder of this section, every unique metric, identified by its clustered name, will be described in detail.

Given that some of these concepts are rather abstract, and in order to do this in a more concrete manner, in some cases, the example scenario of a bank loan approval will be used.

Firstly, some general notation, which will be used in the remainder of the section for the metrics and the bank loan scenario:

- $X \in R^d$ : quantified features of the applicant (e.g., education, work experience, college GPA, income, etc.). These

**Table 6** Number of times that a metric was referred or mentioned in the articles (left) and the Number of times that a specific metric was referred or mentioned in the articles based on the clustering of similar definitions into a single metric (right). This clustering of similar definitions was established by comparing many references from different authors and by analyzing the definitions and formulas of multiple metrics that were found to be identical. This extensive analysis is presented in this section for each metric in particular

Metric	Times mentioned	Clustered metric	Total
Statistical parity	8	Demographic parity	15
Demographic parity	6		
Statistical distance	1		
Equalized odds	6	Equalized odds	11
Accuracy parity	5		
Equal opportunity	7	Equal opportunity	10
False negative rate balance	1		
False negative rate parity	1		
False negative rates difference	1		
Calibration	3	Predictive parity	9
Predictive parity	2		
False discovery rate difference	1		
False discovery rate parity	1		
False omission rate difference	1		
False omission rate parity	1		
Counterfactual	2	Counterfactuals	5
Counterfactual analysis	1		
Counterfactual comparison	1		
Predictive counterfactual fairness	1		
Disparate impact	4	Disparate impact	5
Proportional parity	1		
False positive rate balance	1	Predictive equality	3
Predictive equality	1		
False positive rate parity	1		
Generalized entropy index	1	Generalized entropy index	2
Theil index	1		
Average odds difference	2		
Error difference	2		
Error ratio	1		
False discovery rate ratio	1		
False negative rates ratio	1		
False omission rate ratio	1		
False positive rates ratio	1		

are the observed characteristics of an individual (variables);

- $A$ : sensitive attribute/protected feature (e.g., sex, race, ethnicity);
- $C = c(X, A) \in \{0, 1\}$ : binary predictor (classifier) (e.g., approved/rejected), which constitutes a decision based on a score  $R := r(x, a) \in [0, 1]$ ; in this case  $c(X, A)$  decides if the person should be given (1) or denied (0) a loan.
- $Y \in \{0, 1\}$ : target variable (binary outcome variable) (e.g., approved/denied);
- $\hat{Y} \in \{0, 1\}$  predicted decision of target variable;
- Assume  $X, A, Y$  are generated from an underlying distribution  $D$  i.e.,  $(X, A, Y) \sim D$ ;
- Denote  $P_a[Y] := P[Y|A = a]$ .

We also briefly describe some concepts that are fundamental to understand the metrics (Fig. 3):

- True positive (TP) is a granted loan to someone who can pay it back;
- False positive (FP) is a granted loan that goes on to default (failed to pay the loan);
- False negative (FN) is a loan denied to someone who can pay it back;

		Predicted condition			
		Positive (PP)	Negative (PN)	Informedness, bookmaker informedness (BM) $= \text{TPR} + \text{TNR} - 1$	Prevalence threshold (PT) $= \frac{\sqrt{\text{TPR} \times \text{FPR}} - \text{FPR}}{\text{TPR} - \text{FPR}}$
Actual condition	Total population $= P + N$				
	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{\text{TP}}{P} = 1 - \text{FNR}$	False negative rate (FNR), miss rate $= \frac{\text{FN}}{P} = 1 - \text{TPR}$
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection	False positive rate (FPR), probability of false alarm, fall-out $= \frac{\text{FP}}{N} = 1 - \text{TNR}$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{\text{TN}}{N} = 1 - \text{FPR}$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{\text{TP}}{\text{PP}} = 1 - \text{FDR}$	False omission rate (FOR) $= \frac{\text{FN}}{\text{PN}} = 1 - \text{NPV}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Negative likelihood ratio (LR-) $= \frac{\text{FNR}}{\text{TNR}}$
	Accuracy (ACC) $= \frac{\text{TP} + \text{TN}}{P + N}$	False discovery rate (FDR) $= \frac{\text{FP}}{\text{PP}} = 1 - \text{PPV}$	Negative predictive value (NPV) $= \frac{\text{TN}}{\text{PN}}$ $= 1 - \text{FOR}$	Markedness (MK), deltaP ( $\Delta p$ ) $= \text{PPV} + \text{NPV} - 1$	Diagnostic odds ratio (DOR) $= \frac{\text{LR}^+}{\text{LR}^-}$
	Balanced accuracy (BA) $= \frac{\text{TPR} + \text{TNR}}{2}$	$F_1$ score $= \frac{2\text{PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$	Fowlkes–Mallows index (FM) $= \sqrt{\text{PPV} \times \text{TPR}}$	Matthews correlation coefficient (MCC) $= \sqrt{\text{TPR} \times \text{TNR} \times \text{PPV} \times \text{NPV}}$ $= \sqrt{\text{FNR} \times \text{FPR} \times \text{FOR} \times \text{FDR}}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}$

Fig. 3 Diagnostic testing diagram. Retrieved from: Wikipedia:[https://en.wikipedia.org/wiki/Precision\\_and\\_recall](https://en.wikipedia.org/wiki/Precision_and_recall)

- True negative (TN) is a loan denied to someone who would default;
- True positive rate (TPR) is the proportion of people who could pay back loans that were actually granted loans;
- False positive rate (FPR) is the proportion of people who would default that were granted loans;
- False negative rate (FNR) is the proportion of people who could pay back loans that were actually denied loans;
- True negative rate (TNR) is the proportion of people who would default that were actually denied loans;

In this situation, both false positives and false negatives are detrimental to both parties, as a false positive costs the lender due to the unpaid loan debt and a false negative costs the borrower due to the financial damage caused by loan default.

A false negative results in financial loss for the bank as it prevents the collection of interest on a repayable loan, while it also imposes a cost on the borrower by denying them access to credit that is rightfully theirs.

Next, we detail each of the metrics identified in the literature.

### 3.3.1 Demographic parity (DP)

Demographic parity or statistical parity is the same concept according to the literature [49]. Statistical distance, in this case, refers to the difference between the probability of a

prediction being positive between two different groups, the same as both demographic and statistical parity. This metric is also known as *independence* in some cases.

The goal of demographic parity is to make the selection of each segment of a group's probabilities as equal as possible, indicating that the model's predictions are not biased with respect to demographic attributes. Demographic parity is the property that the demographics of those receiving positive (or negative) outcomes are identical to the demographics of the population as a whole. Demographic parity speaks to group fairness rather than individual fairness and appears desirable, as it equalizes outcomes across protected and non-protected groups [10, 50].

The equation for demographic parity is relatively straightforward and is typically expressed as follows:

For all  $a, b \in A$ :

$$DP := P(Y = y|A = a) = P(Y = y|A = b) \quad (1)$$

In this equation:

$P(Y = 1|A = a)$  represents the probability that the actual outcome of the target variable ( $Y$ ) is positive (e.g., approval for a loan) given that the individual belongs to demographic group  $a$  (e.g., a protected group, like race = black).

$P(Y = 1|A = b)$  represents the probability that the actual outcome of the target variable is positive given that the individual belongs to demographic group  $b$  (e.g., a non-protected group, like race = white).

### 3.3.2 Equalized odds (EqO)

Equalized odds is a bias mitigation technique combination of both equal opportunity and predictive equality, where a classifier has equality of odds when the protected and unprotected groups achieve equality of TPR (equal opportunity) and FPR (predictive equality) across the two groups [51].

Equalized Odds now considers conditional expectations with respect to positive and negative labels, i.e.,  $Y = 0$  and  $Y = 1$ . In order to meet this criterion, the outcomes of the subset of records belonging to the positive and negative classes must be equal. In certain articles, it is also denoted as false positive rate parity [51].

As previously stated, true positive parity is occasionally referred to as equality of opportunity because it mandates that the entire population, irrespective of the dominant group, is afforded the chance to benefit from the decision ( $Y = 1$ ) (See Eq. (5)).

Likewise, false positive rate parity, sometimes referred to as predictive equality, is described in Eq. (14).

Mathematically, it is equivalent to the conjunction of conditions for false positive error rate balance and false negative error rate balance definitions given above. In this instance, this implies that the probability of an applicant with an actual good credit score being correctly assigned a good predicted credit score and the probability of an applicant with an actual bad credit score being incorrectly assigned a good predicted credit score should both be the same for male and female applicants, for example [52].

$$\begin{aligned} EqO := P(\hat{Y} = 1|Y = y, A = a) = \\ P(\hat{Y} = 1|Y = y, A = b), y \in 0, 1 \end{aligned} \quad (2)$$

The inclusion of false positive rates takes into consideration the varying costs associated with misclassification across different populations. In instances where a model is utilized to forecast an unfavorable result, such as the likelihood of recidivism, and this outcome disproportionately impacts individuals from minority communities, the presence of false positive predictions can be seen as a manifestation of pre-existing disparities in outcomes between minority and majority populations. The concept of equalized odds aims to ensure that the predictive accuracy of models is uniformly high across all demographic groups. This approach penalizes models that exhibit strong performance solely on the majority groups.

A weaker variant of the equalized odds is referred to as Accuracy parity. The achievement of accuracy parity is realized when the accuracies within subgroups, which are determined by dividing the total number of successfully classified examples by the overall number of examples, exhibit little disparity. One drawback of this less robust concept is

the potential trade-off between the false positive rate of one group and the false negative rate of another group.

Accuracy parity requires that the classifier guarantees the same accuracy in different sensitive attribute groups.

Accuracy (ACC) is defined as [10]:

$$ACC_a = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

So, accuracy parity can be represented by:

$$ACCP := P(\hat{Y} = Y|A = a) = P(\hat{Y} = Y|A = b) \quad (4)$$

### 3.3.3 Equal opportunity (EO)

Equal opportunity is a state of fairness characterized by a fair treatment of individuals, where they are afforded identical opportunities and rights. The principle asserts that the representation of any subgroup within a larger population (such as gender or ethnicity) should be determined based on the relative size of that subgroup in the overall population. This determination should be made without any hindrance from artificial obstacles, prejudices, or preferences, unless specific justifications can be provided for certain distinctions [10, 18].

Equal opportunity is a relaxed version of equalized odds that only considers conditional expectations with respect to positive labels (i.e.,  $Y = 1$ ). This metric requires equal outcomes only within the subset of records belonging to the positive class and is defined in some cases as equal true positive rates (true positive parity or balance).

In the loaning example, equal opportunity requires that the individuals in group  $a$  who are qualified to be given a loan are just as likely to be chosen as individuals in group  $b$  who are qualified to be given a loan. However, by not considering whether false positive rates are equivalent across groups, equal opportunity does not capture the costs of misclassification disparities [16].

EO calculates the ratio of true positives to positive examples in the dataset  $TPR = TP/P$ , conditioned on a protected attribute.

Likewise, false negative error rate balance measures the probability of a subject in a positive class to have a negative predictive value being the same across the groups. So, equal opportunity, false negative rate parity, false negative rate difference and false negative rate balance are similar.

A common denominator that emerged in the systematic literature review is that different articles presented the same metrics with different names. A conclusion that was taken is that every time an article addressed *difference*, *parity* and *balance* they were describing or referring to the same metric and trying to achieve the same conclusions based on slightly different equations/calculations. So in this case, false nega-



tive rate parity, difference and balance are exactly the same metric, justifying its clustering into a same ethic metric, this being equal opportunity [16, 52, 53].

For all  $a, b \in A$ :

$$\begin{aligned} \text{EO} &:= P(\hat{Y} = 1|Y = 1, A = a) \\ &= P(\hat{Y} = 1|Y = 1, A = b) \end{aligned} \quad (5)$$

FNR difference measures the equality (or lack thereof) of false negative rates between groups. In practice, this metric (false negative rate parity) is implemented as a difference between the metric value for group  $a$  and group  $b$  [52].

$$\begin{aligned} \text{FNRD} &:= P(\hat{Y} = 0|Y = 1, A = a) \\ &= P(\hat{Y} = 0|Y = 1, A = b) \end{aligned} \quad (6)$$

The condition for a prediction to be considered accurate is satisfied when the false negative rates (FNRs) for both the protected and unprotected categories are equal. In other words, the FNR indicates the probability that an individual in the positive class will receive a negative prediction, specifically, in this context, the likelihood that an individual who has the ability to pay back a loan will be rejected one.

In this example, this implies that the probability of an applicant with an actual good credit score being incorrectly assigned a bad predicted credit score should be the same for both male and female applicants, for instance (or for both white and black, etc.):  $P(\hat{Y} = 0|Y = 1, A = a) = P(\hat{Y} = 0|Y = 1, A = b)$ . Mathematically, a classifier with equal FNRs will also have equal TPR:  $P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = b)$  [18, 52].

For that matter, when evaluating AI based on equal opportunity, both TPR balance or FNR balance are acceptable.

### 3.3.4 Predictive parity (PrP)

Predictive parity, calibration, or false omission rate difference (or parity) can also be described as *Sufficiency* [54] in some instances.

The predictive value parity is satisfied when both positive predictive value parity (PPV-parity) and negative predictive value parity (NPV-parity) are satisfied, for both protected and unprotected groups.

PPV is the probability of a subject with positive predictive value to truly belong to the positive class. In the bank loan scenario used as example, PPV is the proportion of granted loans that were paid back, and the NPV is the proportion of rejected loans that were denied to someone who could not pay the loan [52].

PPV-parity equalizes the chance of success, given a positive prediction. For example, in the bank's example, PPV-

parity requires credit score rates to be equal across groups (of admitted individuals).

So, for all  $a, b \in A$ :

$$\begin{aligned} \text{PrP} &:= P(Y = y|\hat{Y} = \hat{y}, A = a) = \\ &P(Y = y|\hat{Y} = \hat{y}, A = b), y \in 0, 1 \end{aligned} \quad (7)$$

This is equivalent to satisfying both:

$$\begin{aligned} \text{PPVP} &:= P(Y = 1|\hat{Y} = 1, A = a) \\ &= P(Y = 1|\hat{Y} = 1, A = b) \end{aligned} \quad (8)$$

$$\begin{aligned} \text{NPVP} &:= P(Y = 0|\hat{Y} = 0, A = a) \\ &= P(Y = 0|\hat{Y} = 0, A = b) \end{aligned} \quad (9)$$

In this example, this implies that, the score returned from a prediction (used to determine the individuals's eligibility for the loan) for an individual, should reflect the person's real capability of paying for the loan.

In other words, for example, for both male and female applicants, the probability that an applicant with a good predicted credit score will actually have a good credit score should be the same [52].

Furthermore, mathematically, a classifier with equal PPVs will also have equal false discovery rates (FDRs):

$$\begin{aligned} \text{FDRP} &:= P(Y = 0|\hat{Y} = 1, A = a) = \\ &P(Y = 0|\hat{Y} = 1, A = b). \end{aligned} \quad (10)$$

An additional equation to prove false omission rate parity is the other side of predictive parity:

$$\begin{aligned} \text{FORP} &:= P(Y = 1|\hat{Y} = 0, A = a) = \\ &P(Y = 1|\hat{Y} = 0, A = b) \end{aligned} \quad (11)$$

We conclude that when trying to measure the predictive parity, the positive predictive parity, the negative predictive parity, the false omission rate parity / difference and the false discovery rate parity are all implicitly or explicitly obtained and measured.

### 3.3.5 Counterfactuals (CF)

Kusner et al. [55] introduced the notion of Counterfactual fairness, which is a form of fairness developed from Pearl's causal model. In this framework, the fairness of a prediction made by the model for an individual is determined by whether it remains consistent when the individual is hypothetically assigned to a different demographic group [20].

The authors propose to employ counterfactuals and define a decision-making process counterfactually fair if, for any individual, the outcome does not change in the counterfactual scenario where the sensitive attributes are changed.

To measure this, they make explicit assumptions about the causal relationships in the data. One way for a predictor to be counterfactually fair is if it is a function of only non-descendants of the sensitive attribute, so this will be different depending on the chosen causal model [20, 55].

In other words, the predictor  $\hat{Y}$  is counterfactually fair if under any context  $Y = y$  and  $A = a$ . The mathematical formulation reads:

$$CF := P(\hat{Y}_{A \leftarrow a} = y | X = x, A = a) = P(\hat{Y}_{A \leftarrow b} = y | X = x, A = a) \quad (12)$$

That is, taken a random individual with sensitive attribute  $A = a$  and other features  $X = x$  and the same individual if she had  $A = b$ , they should have the same chance of being accepted. The symbol  $\hat{Y}_{A \leftarrow a}$  represents the counterfactual random variable  $\hat{Y}$  in the scenario where the sensitive attribute  $A$  is fixed to  $A = a$ . The conditioning on  $A = a$ ,  $X = x$  means that this requirement is at the individual level, in that is conditioning on all the variables identifying a single observation [55].

### 3.3.6 Disparate impact (DI)

Disparate impact also known as proportional parity or even adverse impact commonly refers to the measurement of unintentional discriminatory practice. It is a quantitative measure of the adverse treatment of protected classes that compares the positive outcome of one group versus another [56].

Disparate impact can also be part of the demographic parity for some authors [10, 20].

DI is then not a difference or a parity, but the ratio of predictions for a favorable outcome in a binary classification task between members of the unprivileged group  $a$  to those of the privileged group  $b$  [10]. For all  $a, b \in A$ :

$$DI := \frac{P(\hat{Y} = 1 | A = a)}{P(\hat{Y} = 1 | A = b)} \quad (13)$$

In the bank loan scenario, it can be the ratio of the prediction for being approved for a loan between female and male, for instance. Disparate impact is a one-side reformulation of this condition, where 80% disparity is an agreed upon tolerance decided in the legal arena. For example, if the model predictions grant loans to 60% of men (group  $b$ ) and 50% to women (group  $a$ ), then  $DI = 0.5/0.6 = 0.8$ , which indicates a positive bias and an adverse impact on the other group represented by  $a$ .

Values less than 1 indicate that group  $b$  has a higher proportion of predicted positive outcomes than group  $a$ . This is referred to as positive bias.

A value of 1 indicates demographic parity.

Values greater than 1 indicate that group  $a$  has a higher proportion of predicted positive outcomes than group  $b$ . This is referred to as negative bias.

### 3.3.7 Predictive equality (PrE)

Predictive equality also referred to as false positive error rate balance is defined as the situation when accuracy of decisions is equal across two groups, as measured by false positive rate (FPR) [53].

A classifier satisfies this condition if subjects in the protected and unprotected groups have equal FPR, as indicated by the fulfillment of the following equation. For all  $a, b \in A$ :

$$PrE := P(\hat{Y} = 1 | Y = 0, A = a) = P(\hat{Y} = 1 | Y = 0, A = b) \quad (14)$$

Mathematically, if a classifier has equal FPR for both groups, it will also have equal TNR, satisfying the equation:

$$TNR := P(\hat{Y} = 0 | Y = 0, A = a) = P(\hat{Y} = 0 | Y = 0, A = b) \quad (15)$$

### 3.3.8 Generalized entropy index (GE)

Generalized entropy index is proposed as a unified measure of individual and group fairness by Speicher et al. [57]. It is a measure of inequality at a group or individual level with respect to the fairness of the algorithmic outcome [58].

This class of inequality indexes is based on the concept of *entropy*. In thermodynamics, entropy is a measure of disorder. When applied to income distributions, entropy (disorder) has the meaning of deviations from perfect equality [59].

Generalized Entropy Index measures the inequality between all users with respect to how fair they are treated by the algorithm. Entropy-based metrics such as Generalized Entropy Index are a family of inequality indices that can be used to measure fairness at both group level and individual level.

The Theil index is the most commonly used flavor of GEI. It can be considered a measure of the inequality between all individuals with respect to how fair the outcome of the algorithm is [57–59].

The GE index is defined as follows [57–60]:

$$GE(\alpha) = \frac{1}{n\alpha(\alpha - 1)} \sum_{i=1}^n \left[ \left( \frac{y_i}{y} \right)^\alpha - 1 \right] \quad (16)$$

The values of the  $GE(\alpha)$  index vary between 0 and  $\infty$ , with 0 representing an equal income distribution and higher values representing higher levels of income inequality. The  $GE(\alpha)$  index as shown in Eq. (16) defines a class because it

assumes different forms depending upon the value assigned to the parameter  $\alpha$ , which is a weight given to inequalities in different parts of the income distribution [60]. A positive  $\alpha$  captures the sensitivity of the  $GE$  index to a specific part of the income distribution. With positive and large  $\alpha$ , the index  $GE$  will be more sensitive to what happens in the upper tail of the income distribution. The less positive the parameter  $\alpha$  is, the more sensitive the index is to inequalities at the bottom of the income distribution while the more positive the parameter  $\alpha$  is, the more sensitive the index is to inequalities at the top [59, 60].

In principle, the parameter  $\alpha$  can take any real values from  $-\infty$  to  $\infty$ . However, from a practical point of view,  $\alpha$  is normally chosen to be positive. This is because, for  $\alpha < 0$ , this class of indices is undefined if there are zero incomes.  $GE(0)$  is referred to as the mean logarithmic deviation, which is defined as follows [60]:

$$GE(0) = -\frac{1}{n} \sum_{i=1}^n \ln \left( \frac{y_i}{\bar{y}} \right) \quad (17)$$

$GE(1)$  is known as the Theil inequality index, named after the author who devised it in 1967 [60]. It is calculated as the generalized entropy of benefit for all individuals in the dataset, with  $\alpha = 1$ . Thus, it measures the inequality in benefit allocation for individuals [10].

The Theil index is defined as follows:

$$GE(1) = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{\bar{y}} \ln \left( \frac{y_i}{\bar{y}} \right) \quad (18)$$

### 3.3.9 Average odds difference (AOD)

Average odds difference is a fairness metric used to assess the difference in predictive performance between two groups or populations (between unprivileged and privileged groups) in terms of both false positive rates and true positive rates. It focuses on measuring the balance of the prediction outcomes among different groups [10, 61].

So, the average odds denote the average difference in FPR and TPR for groups  $a$  and  $b$  [18], defined by the following equations.

For all  $a, b \in A$ :

$$AOD := \frac{1}{2} [(FPR_{A=a} - FPR_{A=b}) + (TPR_{A=b} - TPR_{A=a})] \quad (19)$$

This can be translated to:

$$AOD := \frac{1}{2} [(P(\hat{Y} = 1|Y = 0, A = a) - P(\hat{Y} = 1|Y = 0, A = b)) +$$

$$(P(\hat{Y} = 1|Y = 1, A = a) - P(\hat{Y} = 1|Y = 1, A = b))] \quad (20)$$

The ideal value of this metric is 0. A negative value (value  $< 0$ ) implies that the model produces higher false positive rates and true positive rates for the privileged group ( $b$ ), and a positive value (value  $> 0$ ) indicates a higher benefit for the unprivileged group ( $a$ ), both suggesting potential disparities in predictive performance [10].

### 3.3.10 Error difference (ErrD)

Error difference is a metric that is intended to measure the difference between the false positive and false negative rates of unprivileged group ( $a$ ) and the false positive and false negative rates of privileged group ( $b$ ).

The goal is to understand the proportion of missed predictions based on the real value. This metric can be satisfied by the following equation:

For  $a, b \in A$ :

$$ErrD := \frac{FP_a + FN_a}{N_b + N_a} - \frac{FP_b + FN_b}{N_b + N_a} \quad (21)$$

This is equivalent to:

$$ErrD := P(\hat{Y} \neq Y, A = a) - P(\hat{Y} \neq Y, A = b) \quad (22)$$

The ideal number is 0, meaning that both false positive and negative rates for both privileged and unprivileged groups are balanced. A negative value means there is a discrepancy between groups' predictions, meaning there may be potential bias toward the privileged group, since group  $b$  presents a higher false positive and negative rate than group  $a$ . In relation to that, a positive value shows potential bias toward the unprivileged group  $a$ .

### 3.3.11 Error ratio (ErrR)

Likewise, error ratio is a metric with the same intention of the error difference, but now trying to look to the comparison of both groups to try to understand how much of one group is present in the other group, or, in other words, how two distinct groups are related.

The metric is defined by the following equation:

For  $a, b \in A$ :

$$ErrR := \frac{\frac{FP_a + FN_a}{N_b + N_a}}{\frac{FP_b + FN_b}{N_b + N_a}} \quad (23)$$

This is equivalent to:

$$\text{ErrR} := \frac{P(\hat{Y} \neq Y, A = a)}{P(\hat{Y} \neq Y, A = b)} \quad (24)$$

The ideal number is 1, which means that both false positive and negative rates for privileged and unprivileged groups are balanced. A value close to zero indicates that there is a large discrepancy between groups, meaning there are disparities and potential bias toward the privileged group, since group  $b$  presents a higher false positive and negative rate than group  $a$ . In relation to that, a positive value greater than 1 shows a potential bias toward the unprivileged group  $a$ .

### 3.3.12 False discovery rate ratio (FDRR)

Similar to false discovery rate parity, the false discovery rate ratio is a method of conceptualizing the rate of type I errors in null hypothesis testing when conducting multiple comparisons between two different groups.

This metric is used to assess the proportion of false positive predictions among all positive predictions made by a classification model between two different sensitive attributes. It quantifies the ratio of incorrect positive predictions relative to the total number of positive predictions of sensitive attribute  $a$  and attribute  $b$  and then compares them.

The formula for false discovery rate ratio is relatively straightforward and is typically expressed as follows:

$$\text{FDRR} := \frac{P(Y = 0 | \hat{Y} = 1, A = a)}{P(Y = 0 | \hat{Y} = 1, A = b)} \quad (25)$$

### 3.3.13 False negative rate ratio (FNRR)

Similar to false negative rate parity, the false negative rate ratio is used to assess the equality (or lack thereof) of the false negative rates across groups. It quantifies the ratio of incorrect negative predictions relative to the total number of positive cases for the sensitive group  $a$  and the group  $b$  and then compares them.

The equation for false negative rate ratio is typically expressed as follows:

$$\text{FNRR} := \frac{P(\hat{Y} = 0 | Y = 1, A = a)}{P(\hat{Y} = 0 | Y = 1, A = b)} \quad (26)$$

### 3.3.14 False omission rate ratio (FORR)

Similar to false omission rate parity, the false omission rate ratio is used to assess the equality (or lack thereof) between groups of the rate of inaccurate “negative” predictions by the model.

The formula for false negative rate ratio is typically expressed as follows:

$$\text{FORR} := \frac{P(Y = 1 | \hat{Y} = 0, A = a)}{P(Y = 1 | \hat{Y} = 0, A = b)} \quad (27)$$

### 3.3.15 False positive rate ratio (FPRR)

Similar to false positive rate parity, this metric is used to assess the equality (or lack thereof) of the false positive rates across groups. It quantifies the ratio of incorrect positive predictions relative to the total number of negative cases, for both unprivileged group  $a$  and privileged group  $b$ , comparing them.

The formula for false positive rate ratio is typically expressed as follows:

$$\text{FPRR} := \frac{P(\hat{Y} = 1 | Y = 0, A = a)}{P(\hat{Y} = 1 | Y = 0, A = b)} \quad (28)$$

## 3.4 Additional information

Some of the articles analyzed mentioned performance metrics such as *accuracy* (12), *precision* (5), *AUC—Area Under ROC Curve* (3), *recall* (3), *RMSE—Root Mean Squared Error* (1), *MSE—Mean Squared Error* (1), *F1-score* (1), *Mann–Whitney U test* (1), among others. These are, however, in our view a different kind of metric, which quantifies the overall predictive quality of a ML model. While these metrics might be relevant to assess the level of ethical compliance of a system, be relevant for ethical requirements such “Technical Robustness and Safety” and would surely be useful to create ethics metrics, they cannot be exclusively considered ethics metrics *per se*. For instance, accuracy can evaluate the predictive quality of a model but not whether the model is ethical; a model can have good overall/average accuracy but poor accuracy for specific groups (discrimination). Not only that, but these performance metrics obtained through the full-text assessment were mostly referenced to evaluate a model in terms of performance but not in terms of ethical compliance. Additionally, although these performance metrics are relevant and shouldn’t be completely overlooked and discarded, they were not discussed like the previous ones since they have received plenty of attention in non-ethical contexts already.

For this reason, we consider their future use in building novel ethics metrics, but we did not consider them in this literature review with the same level of detail as the previous ones.

Another relevant disclaimer about this systematic literature review is that we found several papers that, when addressing the ethical compliance of AI, presented some



solutions/tools or methods for ethical analysis, bias mitigation, or model performance enhancement. Such tools and metrics that were identified in the literature (e.g., SHAP, LIME, conformity assessment, BLEU score, average popularity GAP, transparency score, reporting rate, etc.) are undoubtedly relevant in AI Ethics. However, while they can, to some extent, facilitate the implementation of efforts toward Ethical AI, they are not objective metrics *per se*, as discussed previously, and are therefore outside of the scope of this literature review. While mentioned here, they are not given the same level of detail as the identified metrics.

For instance, Ruf and Detyniecki [25] present two tools that can help operationalize AI fairness. The fairness compass helps choose what fairness means for each specific situation through a fairness tree. The Fairness Library helps choosing the best way to make AI fair by providing a series of algorithms for bias mitigation. This specific paper is an example of one that presents a relevant tool, but does not put forward any specific metric.

The following sections discuss the main findings and limitations of this work and the key conclusions.

## 4 Discussion

This work started with the goal and motivation to find a wide range of objective ethics metrics, catering to, if possible, all of the majority of the widely accepted principles. This would make it possible for Data Scientists to have a list of observable metrics, that they could easily implement and integrate into any Data Science pipeline in order to assess its alignment with ethical principles and guidelines.

However, it results from this literature review that the subject of ethics in the area of AI is still relatively immature. While it has gained significant theoretical attention as one of the prominent subjects of discussion, practical solutions in this area are still underdeveloped. Maybe more interdisciplinary efforts that include, aside from ethics experts, specialists such as Computer Scientists, Data Scientists, ML Engineers, or AI Architects could be beneficial.

In face of the lack of practical solutions, the prevailing discourse predominantly revolves around the future trajectory of AI and the imperative of ethical considerations, yet only a minority of researchers actively engage in developing viable, measurable and practical solutions to address this challenge.

While several of the articles identified analyzed the challenges of AI, the ethical needs for AI, or how AI is impacting society, these were mostly theoretical works. Thus, it was challenging to draw specific conclusions or identify actual metrics.

Furthermore, it is generally acknowledged that all ethical principles are relevant when evaluating a Data Science pipeline, although some variations may occur between fields.

For instance, while privacy is always important, it might be more important in healthcare or education than in certain industrial settings. However, the lack of ethics metrics in 6 out of 7 ethical principles can be seen as potentially hindering the progress toward a safe, fair and all-encompassing ethical AI. The reason for this unbalance can be explained, from our perception, by the relative facility that there is in assessing fairness by following a distributive and comparative approach. Generally, by analyzing how data (e.g., raw data, processed data, model outputs) are distributed across sensitive features.

We must thus be aware that the field of AI Ethics is nowadays very much biased toward Fairness, with a significant disregard for other principles. This is, in our opinion, both a current handicap of the field but also a major research opportunity.

Interestingly, a superficial analysis of the literature might appear to indicate that there are many more metrics than those identified in this literature review. However, this happens as many of the existing metrics overlap significantly or represent the same concept, albeit with a different denomination. A relevant contribution of this work was to reduce such groups of equivalent metrics into a single one, contributing to a clarification of the terminology and of the metrics available. Specifically, we reduced the original set of 33 metrics to 15, which still represent the same concepts being measured.

Aside from this, this SLR, which is, to the best of our knowledge, the first one with these goals, also systematized the existing knowledge regarding objective metrics that can be used to quantify the ethical alignment of a Data Science pipeline. As such, Data Scientists can use it to select the most appropriate metrics to monitor their systems, in a transparent and standardized way. Unfortunately, the existing metrics in the literature only allow to assess the system in terms of fairness.

We plan on bridging this gap in future work by proposing and implementing, in the form of a software library, a group of metrics for each of the identified ethical principles. This library will connect to different batch and streaming data sources and allow for the monitoring of the associated ethical principles out of the shelf. These metrics will be validated in different Data Science pipelines across different domains in an attempt to evaluate their truthfulness and usefulness. While we expected this implementation and development work to follow directly from this work, the current state of affairs demands we first devise and implement the lacking metrics.

### 4.1 Main findings

This systematic review of the literature on the topic of objective metrics to assess AI ethics identified 38 papers, which were considered for full-text analysis. These were organized

according to seven main ethical principles (defined as per the ethics guidelines for trustworthy AI) based on the subjects addressed.

According to the results of the analysis, 24 articles addressed Diversity, Non-Discrimination and Fairness, 14 articles discussed Transparency, 7 articles focused on Accountability, 6 addressed Privacy and Data Governance, 6 were associated with Societal and Environmental Well-being, 3 articles related to Technical Robustness and Safety, and 2 addressed Human Agency and oversight.

Out of these 38 studies, only 12 presented actual objective metrics, all of which related to Diversity, Non-Discrimination and Fairness, which can be assigned to the well-known category of Fairness Metrics.

Out of these metrics, the ones that were most frequently mentioned were demographic parity, equal opportunity, predictive parity and equalized odds.

This research exposes several key aspects:

- There is a significant gap between theoretical research and practice, or practical solutions. While the need for metrics related to different principles is often mentioned, the proposal of actual metrics that can be operationalized is scarce;
- The research devoted to each principle is highly unbalanced. Some principles are much more prevalent than others (whether in theory or practice). Moreover, all the metrics found are related to the principle of Fairness, which means that the adherence to the remaining principles cannot be objectively evaluated in current AI systems;
- The apparent wealth of metrics that are mentioned in the literature is not actually so rich, as many of the proposed metrics are actually slight variations (sometimes with only a change in the denomination) or sometimes measuring the exact same concept in a slightly different way;
- Similarly, many ethical principles have been proposed in the literature. However, many of them overlap or are closely related. Although some authors argue that the seven AI-HLEG principles around which this literature review is organized are not maximally representative [5, 62–64], we find that all the metrics identified and the related principles can, in fact, be grouped into those seven main principles.

## 5 Conclusion

In this systematic literature review, a total of 38 papers, from a pool of 66 candidates, were examined. These papers were categorized according to seven ethical principles. There were two articles that focused on the topic of Human

Agency and Oversight, three articles that addressed Technical Robustness and Safety, six articles that explored Privacy and Data Governance, fourteen articles that examined Transparency, twenty-two articles that explored into Diversity, Non-Discrimination and Fairness, six articles that discussed Societal and Environmental Well-being and seven articles that were related to Accountability.

The bulk of studies consisted of academic publications that articulated the necessity of addressing ethical concerns in the field of AI and proposed potential theoretical frameworks for achieving this objective. However, most studies did not offer concrete, practical answers or demonstrate real-world applications.

Among the articles subjected to comprehensive textual analysis, only 12 articles were found to incorporate at least one discernible and practical ethics metric. Notably, these metrics were predominantly centered around fairness and pertained to the ethical principle of Diversity, Non-Discrimination and Fairness.

The most frequent metrics were Demographic Parity (15 times), Equalized Odds (11 times), Equal Opportunity (10 times), Predictive Parity (9 times), Counterfactuals (5 times), Disparate Impact (5 times) and Predictive Equality (3 times).

The key conclusion drawn from this study is that there is a severe lack of practical solutions for the ethical assessment of Data Science pipelines in six out of the seven ethical principles considered. Thus, there is both the need and the scientific opportunity to develop such embracing metrics. Nonetheless, the existing metrics were systematized, and the generated knowledge can be used by Data Scientists to integrate standardized, transparent, and objective ethics metrics into their pipelines, ensuring that their applications are observable in terms of their alignment with ethical standards and guidelines.

**Acknowledgements** This work has been supported by the European Union under the Next Generation EU, through a grant of the Portuguese Republic's Recovery and Resilience Plan (PRR) Partnership Agreement, within the scope of the project PRODUTECH R3 —“Agenda Mobilizadora da Fileira das Tecnologias de Produção para a Reindustrialização,” Total project investment: 166.988.013,71 Euros; Total Grant: 97.111.730,27 Euros. The work of Guilherme Palumbo was funded by a research grant from FCT—Fundação para a Ciência e Tecnologia under Grant Agreement No. [UIDP/04728/2020].

**Author Contributions** GP did the acquisition, analysis and interpretation of results. All authors wrote the main manuscript text. All authors reviewed the manuscript. All authors approved the version to be published.

**Funding** Open access funding provided by FCTIFCCN (b-on). This work has been supported by the European Union under the Next Generation EU, through a grant of the Portuguese Republic's Recovery and Resilience Plan (PRR) Partnership Agreement, within the scope of the project PRODUTECH R3—“Agenda Mobilizadora da Fileira das Tecnologias de Produção para a Reindustrialização,” Total project investment: 166.988.013,71 Euros; Total Grant: 97.111.730,27 Euros.

The work of Guilherme Palumbo was funded by a research grant from FCT—Fundação para a Ciência e Tecnologia under Grant Agreement No. [UIDP/04728/2020].

**Data Availability** Not applicable.

**Code Availability** Not applicable.

## Declarations

**Conflict of interest** Author Guilherme Palumbo has received research support from FCT—Fundação para a Ciência e Tecnologia under Grant Agreement No[UIDP/04728/2020]. Author Davide Carneiro and Victor Alves declare they have no financial interests.

**Ethical approval** Not applicable.

**Consent to participate** All authors consent to participate.

**Consent for publication** All authors consent for publication.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Gichoya, J.W., Banerjee, I., Bhimireddy, A.R., Burns, J.L., Celi, L.A., Chen, L.-C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S.-C., et al.: AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit. Health* **4**(6), 406–414 (2022)
- Hunkenschroer, A.L., Luetge, C.: Ethics of AI-enabled recruiting and selection: a review and research agenda. *J. Bus. Ethics* **178**, 977–1007 (2022). <https://doi.org/10.1007/s10551-022-05049-6>
- Barrett, C., Boyd, B., Bursztein, E., Carlini, N., Chen, B., Choi, J., Chowdhury, A.R., Christodorescu, M., Datta, A., Feizi, S., Fisher, K., Hashimoto, T., Hendrycks, D., Jha, S., Kang, D., Kerschbaum, F., Mitchell, E., Mitchell, J., Ramzan, Z., Shams, K., Song, D., Taly, A., Yang, D.: Identifying and mitigating the security risks of generative AI. *Found. Trends Privacy Secur.* **6**(1), 1–52 (2023). <https://doi.org/10.1561/33000000041>
- CDEI: The roadmap to an effective AI assurance ecosystem - GOV.UK (2021)
- BridgeAI, I.U.: Report on the core principles and opportunities for responsible and trustworthy AI (2023)
- European Commission: Ethics guidelines for trustworthy AI | Shaping Europe's digital future (2019). <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- 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., Zhang, Y.: AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias (2018). <https://arxiv.org/abs/1810.01943>
- Siddaway, A.P., Wood, A.M., Hedges, L.V.: How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annu. Rev. Psychol.* **70**, 747–770 (2019). <https://doi.org/10.1146/annurev-psych-010418>
- PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses. <http://prisma-statement.org/>
- Kaul, A., Soofastaei, A.: Advanced analytics for ethical considerations in mining industry. In: *Advanced Analytics in Mining Engineering: Leverage Advanced Analytics in Mining Industry to Make Better Business Decisions*, pp. 55–80. (2022) [https://doi.org/10.1007/978-3-030-91589-6\\_3/COVER](https://doi.org/10.1007/978-3-030-91589-6_3/COVER)
- Kasirzadeh, A.: Algorithmic fairness and structural injustice: insights from feminist political philosophy. In: *AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 349–356. (2022) <https://doi.org/10.1145/3514094.3534188>
- Zafar, S., Mahjoub, H., Mehta, N., Domalpally, A., Channa, R.: Artificial intelligence algorithms in diabetic retinopathy screening. *Curr. Diabet. Rep.* **22**, 267–274 (2022). <https://doi.org/10.1007/S11892-022-01467-Y/METRICS>
- Bae, A., Xu, S.: Discovering and understanding algorithmic biases in autonomous pedestrian trajectory predictions. In: *SenSys 2022 - Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pp. 1155–1161. (2022) <https://doi.org/10.1145/3560905.3568433>
- Kasirzadeh, A., Clifford, D.: Fairness and data protection impact assessments. In: *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 146–153. (2021) <https://doi.org/10.1145/3461702.3462528>
- Marshall, R., Pardo, A., Smith, D., Watson, T.: Implementing next generation privacy and ethics research in education technology. *Br. J. Educ. Technol.* **53**, 737–755 (2022). <https://doi.org/10.1111/BJET.13224>
- Cortés, E.C., Rajtmajer, S., Ghosh, D.: Locality of technical objects and the role of structural interventions for systemic change. In: *ACM International Conference Proceeding Series*, pp. 2327–2341. (2022) <https://doi.org/10.1145/3531146.3534646>
- Abedin, B.: Managing the tension between opposing effects of explainability of artificial intelligence: a contingency theory perspective. *Internet Res.* **32**(2), 425–453 (2021). <https://doi.org/10.1145/3479645.3479709>
- Zhang, Q., Liu, J., Zhang, Z., Wen, J., Mao, B., Yao, X.: Mitigating unfairness via evolutionary multiobjective ensemble learning. *IEEE Trans. Evolut. Comput.* **27**, 848–862 (2023). <https://doi.org/10.1109/TEVC.2022.3209544>
- Schedl, M., Rekabsaz, N., Lex, E., Grosz, T., Greif, E.: Multi-perspective and multidisciplinary treatment of fairness in recommender systems research. In: *UMAP2022 - Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 90–94. (2022) <https://doi.org/10.1145/3511047.3536400>
- Goethals, S., Martens, D., Calters, T.: Precof: counterfactual explanations for fairness. *Mach. Learn.* (2023). <https://doi.org/10.1007/S10994-023-06319-8/FIGURES/10>
- Tomalin, M., Byrne, B., Concannon, S., Saunders, D., Ullmann, S.: The practical ethics of bias reduction in machine translation: why domain adaptation is better than data debiasing. *Ethics Inform. Technol.* **23**, 419–433 (2021). <https://doi.org/10.1007/S10676-021-09583-1/TABLES/7>
- Fleisher, W.: What's fair about individual fairness? In: *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 480–490. (2021) <https://doi.org/10.1145/3461702.3462621>

23. Wachter, S., Mittelstadt, B., Russell, C.: Why fairness cannot be automated: bridging the gap between EU non-discrimination law and AI. *SSRN Electron. J.* (2020). <https://doi.org/10.2139/SSRN.3547922>
24. Saetra, H.S., Wynsberghe, Bolte, L., Nachid, J.: A framework for evaluating and disclosing the esg related impacts of AI with the sdgs. *Sustainability* **13**, 8503 (2021). <https://doi.org/10.3390/SU13158503>
25. Ruf, B., Detyniecki, M.: A tool bundle for ai fairness in practice. In: *Conference on Human Factors in Computing Systems - Proceedings* (2022). <https://doi.org/10.1145/3491101.3519878>
26. Raji, I.D., Buolamwini, J.: Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In: *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435. (2019) <https://doi.org/10.1145/3306618.3314244>
27. Krijger, J.: Enter the metrics: critical theory and organizational operationalization of AI ethics. *AI Soc.* **37**, 1427–1437 (2022). <https://doi.org/10.1007/S00146-021-01256-3/METRICS>
28. Wylde, V., Prakash, E., Hewage, C., Platts, J.: Ethical challenges in the use of digital technologies: Ai and big data. In: *Advanced Sciences and Technologies for Security Applications*, pp. 33–58. (2023) [https://doi.org/10.1007/978-3-031-09691-4\\_3/COVER](https://doi.org/10.1007/978-3-031-09691-4_3/COVER)
29. Sahu, S., Singh, S.K.: Ethics in ai: Collaborative filtering based approach to alleviate strong user biases and prejudices. In: *2019 12th International Conference on Contemporary Computing, IC3 2019* (2019) <https://doi.org/10.1109/IC3.2019.8844875>
30. Keleko, A.T., Kamsu-Foguem, B., Ngouna, R.H., Tongne, A.: Health condition monitoring of a complex hydraulic system using deep neural network and deepshap explainable xai. *Adv. Eng. Softw.* **175**, 103339 (2023). <https://doi.org/10.1016/J.ADVENGSOFT.2022.103339>
31. McCradden, M.D., Joshi, S., Anderson, J.A., Mazwi, M., Goldenberg, A., Shaul, R.Z.: Patient safety and quality improvement: ethical principles for a regulatory approach to bias in healthcare machine learning. *J. Am. Med. Inform. Assoc. JAMIA* **27**, 2024–2027 (2020). <https://doi.org/10.1093/JAMIA/OCAA085>
32. Lee, W.W.: Tools adapted to ethical analysis of data bias. *HKIE Trans. Hong Kong Inst. Eng.* **29**, 200–209 (2022). <https://doi.org/10.33430/V29N3THIE-2022-0037>
33. Minkinen, M., Niukkanen, A., Mäntymäki, M.: What about investors? esg analyses as tools for ethics-based AI auditing. *AI Soc.* **1**, 1–15 (2022). <https://doi.org/10.1007/S00146-022-01415-0/TABLES/5>
34. Edwards, A.: *IEEE P7010-2020 Standard: Use Cases in Ethical Impact on Human Wellbeing Studies* (2020). <https://doi.org/10.13140/RG.2.2.21769.88168>
35. Fasterholdt, I., Naghavi-Behzad, M., Rasmussen, B.S.B., Kjølhede, T., Skjøth, M.M., Hildebrandt, M.G., Kidholm, K.: Value assessment of artificial intelligence in medical imaging: a scoping review. *BMC Med. Imag.* **22**, 1–11 (2022). <https://doi.org/10.1186/S12880-022-00918-Y/FIGURES/2>
36. Etienne, H.: Solving moral dilemmas with AI: how it helps us address the social implications of the covid-19 crisis and enhance human responsibility to tackle meta-dilemmas. *Law Innov. Technol.* **14**, 305–324 (2022). <https://doi.org/10.1080/17579961.2022.2113669>
37. Carlson, K.W.: Safe artificial general intelligence via distributed ledger technology. *Big Data Cogn. Comput.* **3**, 40 (2019). <https://doi.org/10.3390/BDCC3030040>
38. Steele, R.W.: Pediatric quality measures: the leap from process to outcomes. *Curr. Probl. Pediatr. Adolesc. Health Care* **51**, 101065 (2021). <https://doi.org/10.1016/J.CPPEDS.2021.101065>
39. Avelar, P.H.C., Audibert, R.B., Lamb, L.C.: Measuring ethics in ai with ai: A methodology and dataset construction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 13653 LNAI, pp. 370–384. (2021) [https://doi.org/10.1007/978-3-031-21686-2\\_26](https://doi.org/10.1007/978-3-031-21686-2_26)
40. Riley, P.C., Deshpande, S.V., Ince, B.S., Dereje, R., Davidson, C.E., O'Donnell, K.P., Hauck, B.C.: Interpreting chemical detection alarms with live analysis of ml algorithms. In: *Defense + Commercial Sensing*, vol. 23 (2022) <https://doi.org/10.1117/12.2619166>
41. Claire, H., Chang, M.L., Kim, S., Omeiza, D., Brandao, M., Lee, M.K., Jung, M.: Fairness and transparency in human-robot interaction. In: *ACM/IEEE International Conference on Human-Robot Interaction 2022-March*, pp. 1244–1246. (2022) <https://doi.org/10.1109/HRI53351.2022.9889421>
42. Zou, J., Schiebinger, L.: Ensuring that biomedical ai benefits diverse populations. *EBioMedicine* **67**, 103358 (2021). <https://doi.org/10.1016/j.ebiom.2021.103358>
43. Zhao, K., Ma, S., Sun, Z., Liu, X., Zhu, Y., Xu, Y., Wang, X.: Effect of ai-assisted software on inter- and intra-observer variability for the x-ray bone age assessment of preschool children. *BMC Pediatrics* **22**, 644 (2022). <https://doi.org/10.1186/S12887-022-03727-Y>
44. Young, A.T., Xiong, M., Pfau, J., Keiser, M.J., Wei, M.L.: Artificial intelligence in dermatology: a primer. *J. Invest. Dermatol.* **140**, 1504–1512 (2020). <https://doi.org/10.1016/j.jid.2020.02.026>
45. Lawlor, B.: An overview of the 2022 niso plus conference: global conversations/global connections. *Inf. Serv. Use* **42**, 327–376 (2022). <https://doi.org/10.3233/ISU-220178>
46. Antikainen, J., Agbese, M., Alanen, H.-K., Halme, E., Isomäki, H., Jantunen, M., Kemell, K.-K., Rousi, R., Vainio-Pekka, H., Vakkuri, V.: A deployment model to extend ethically aligned ai implementation method eccola. In: *Proceedings of the IEEE International Conference on Requirements Engineering 2021-September*, pp. 230–235. (2021) <https://doi.org/10.1109/REW53955.2021.00043>
47. Germann, C., Marbach, G., Civardi, F., Fucetese, S.F., Fritz, J., Sutter, R., Pfirrmann, C.W.A., Fritz, B.: Deep convolutional neural network-based diagnosis of anterior cruciate ligament tears: performance comparison of homogenous versus heterogeneous knee mri cohorts with different pulse sequence protocols and 1.5-t and 3-t magnetic field strengths. *Invest. Radiol.* **55**, 499–506 (2020). <https://doi.org/10.1097/RLI.0000000000000664>
48. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **54**(6), 1–35 (2021). <https://doi.org/10.1145/3457607>
49. Räz, T.: Group fairness: Independence revisited 10(1145/3442188), 3445876 (2021)
50. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: *Fairness Through Awareness*. (2011) [arXiv:1104.3913](https://arxiv.org/abs/1104.3913)
51. Tang, Z., Zhang, K.: Attainability and optimality: The equalized odds fairness revisited. In: Schölkopf, B., Uhler, C., Zhang, K. (eds.) *Proceedings of the First Conference on Causal Learning and Reasoning*. *Proceedings of Machine Learning Research*, vol. 177, pp. 754–786. (2022). <https://proceedings.mlr.press/v177/tang22a.html>
52. Verma, S., Rubin, J.: Fairness definitions explained. In: *IEEE/ACM International Workshop on Software Fairness*, vol. 18 (2018) <https://doi.org/10.1145/3194770.3194776>
53. Verma, S., Rubin, J.: Fairness definitions explained. In: *Proceedings - International Conference on Software Engineering*, pp. 1–7. (2018) <https://doi.org/10.1145/3194770.3194776>
54. Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I.G., Cosentini, A.C.: A clarification of the nuances in the fairness metrics landscape. *Sci. Rep.* **12**, 1–21 (2022). <https://doi.org/10.1038/s41598-022-07939-1>
55. Kusner, M., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: *Conference on Neural Information Processing Systems*. (2017) [arXiv:1703.06856](https://arxiv.org/abs/1703.06856)



56. Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web. WWW '17, pp. 1171–1180. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2017). <https://doi.org/10.1145/3038912.3052660>
57. Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K.P., Singla, A., Weller, A., Zafar, M.B.: A unified approach to quantifying algorithmic unfairness: Measuring individual and group unfairness via inequality indices, vol. 12 (2018) <https://doi.org/10.1145/3219819.3220046>
58. Ashokan, A., Haas, C.: Fairness metrics and bias mitigation strategies for rating predictions. *Inf. Process. Manag.* **58**, 102646 (2021). <https://doi.org/10.1016/j.ipm.2021.102646>
59. Bellù, L.G., Liberati, P.: Describing income inequality describing income inequality theil index and entropy class indexes. (2006)
60. Sitthiyot, T., Holasut, K.: A simple method for measuring inequality (2020). <https://doi.org/10.1057/s41599-020-0484-6>
61. Zhang, Y., Bellamy, R.K.E., Varshney, K.R.: Joint optimization of ai fairness and utility: a human-centered approach. vol. 10 (1145/3375627), 3375862 (2020)
62. Jobin, A., Ienca, M., Vayena, E.: Artificial intelligence: the global landscape of ethics guidelines (2019)
63. Floridi, L., Cowls, J.: A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review* **1**(1) (2019). <https://hdr.mitpress.mit.edu/pub/l0jsh9d1>
64. Hagendorff, T.: The ethics of ai ethics: an evaluation of guidelines. *Minds Mach.* **30**, 99–120 (2020). <https://doi.org/10.1007/s11023-020-09517-8>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.