applied sciences

MDPI

Review

Discrimination, Bias, Fairness, and Trustworthy AI

Daniel Varona * and Juan Luis Suárez

CulturePlex Laboratory, London, ON N6A 3K6, Canada; jsuarez@uwo.ca
* Correspondence: dvaronac@uwo.ca

Featured Application: To understand the multiple definitions available for the variables "Discrimination", "Bias", "Fairness", and "Trustworthy AI" in the context of the social impact of algorithmic decision-making systems (ADMS), pursuing to reach consensus as working variables for the referred context.

Abstract: In this study, we analyze "Discrimination", "Bias", "Fairness", and "Trustworthiness" as working variables in the context of the social impact of AI. It has been identified that there exists a set of specialized variables, such as security, privacy, responsibility, etc., that are used to operationalize the principles in the Principled AI International Framework. These variables are defined in such a way that they contribute to others of more general scope, for example, the ones studied in this study, in what appears to be a generalization–specialization relationship. Our aim in this study is to comprehend how we can use available notions of bias, discrimination, fairness, and other related variables that will be assured during the software project's lifecycle (security, privacy, responsibility, etc.) when developing trustworthy algorithmic decision-making systems (ADMS). Bias, discrimination, and fairness are mainly approached with an operational interest by the Principled AI International Framework, so we included sources from outside the framework to complement (from a conceptual standpoint) their study and their relationship with each other.

Keywords: discrimination; bias; fairness; trustworthy ADMS; principled AI; social impact of AI; ethics and AI

Citation: Varona, D.; Suárez, J.L. Discrimination, Bias, Fairness, and Trustworthy AI. *Appl. Sci.* **2022**, *12*, 5826. https://doi.org/10.3390/app12125826

Academic Editors: Cristina Portalés Ricart, João M. F. Rodrigues and Pedro J. S. Cardoso

Received: 1 April 2022
Accepted: 6 June 2022
Published: 8 June 2022

1. Introduction

The negative implications associated with the evolution of machine learning (ML), and by extension to artificial intelligent systems (AIS), and the fact that algorithms and models are increasingly complex and less explainable, make it difficult for users/auditors/developers/researchers to identify if AI systems produce outcomes with negative consequences for humans. However, the most disturbing factor in this evolution of AIS is to learn that we keep outsourcing our responsibility for our decisions to the software we use.

There exists a palpable need to audit black-box algorithms, not only from the verification and validation processes staged as part of the software project lifecycle but also from other areas such as policymaking. Both the engineering approach and the stipulation of the regulatory approach need to be incorporated into an integrative mechanism oriented to reducing and mitigating algorithmic decision-making (ADM) systems-produced discriminatory outcomes, analyzed in previous studies [1,2]. The traditional approach conducted to manage discrimination, prejudice or bias, and algorithmic unfairness historically exhibits a reactive character that must be overcome, as criticized in [3]. Additionally, the needed proactive approach must incorporate the determination of possible remedy actions due to discriminatory ADM systems' outcomes. Then, it is not only necessary to coordinate efforts for mobilizing professionals from multiple disciplines of the technical and humanistic fields to reach a better understanding of the problem and its solution but also to standardize their language to achieve a more effective comprehension of the actions that must be deployed to attain trustworthiness in the AI systems.

The Principled AI International Framework could be considered part of the attempt to specify the idea exposed in the previous paragraph in the direction of achieving trustworthy AI as a business model, specifically centered on safeguarding the individual's rights that might be affected by decisions produced by flawed AI solutions. However, previous findings described in [1,2] highlight that there are significant differences in the language used in the regulatory documents forming the referred framework, which can compromise its proper implementation. For instance, the multiple definitions of the objective variables trustworthy AI are intended to be founded upon are listed among the described gaps.

The present study expands on those previous analyses and the Principled AI International Framework itself, where divergences in language, among other difficulties regarding the framework assimilation as a methodological reference for AIS development, were highlighted. We thought it would be pertinent to explore to what extent the variables among the principles and their agency to propitiate trustworthiness in AIS were compromised due to the ambiguities and lack of precision in the use of language. We then conducted an exploratory survey on the notions of discrimination, bias, fairness, and other related variables such as security, privacy, responsibility, etc., aiming to comprehend how they can be articulated in the pursuit of trustworthiness in the context of ADM systems.

2. Related Research

Abhishek and others proposed a framework for trustworthy AI systems [4], providing a data-centric level of abstractions for ethical consideration within the AIS and Data Science contexts that would encompass three levels: data, algorithm, and practice. They defined a set of requirements for trustworthy AIS design. The intended variables coincide with other related studies, although the same cannot be said about some of the proposed definitions, an issue we have already encountered in the principled AI framework. The same happens with others [5,6] and the great majority of the referenced research.

The referenced studies concur with their method, which usually consists of a bibliographic survey determining and defining the variables each paper presents and should be used to build trustworthiness in AI systems. These studies [4–7] distinguish themselves by proposing mechanisms to support the trustworthy AI design by incorporating the variables they each define according to their respective conceptualizations.

These differences might seem trivial; however, in a context totally dependent on the operationalization of the objective variables, this is an aspect that gains major relevance. In fact, it determines the success of the proposed mechanisms and the subsequent achievement of trustworthiness when there is no ambiguity in their implementation.

One of the more critical positions in this approach [8] stresses that the principles show deep political and normative disagreement and highlights that the studied principled framework for AI development lacks common aims and fiduciary duties, proven methods to translate principles into practice, and robust legal and professional accountability mechanisms when comparing it with other frameworks used in fields such as healthcare. This is indeed a very strong criticism of both the principled framework itself and the ways in which it can be operationalized in real contexts of AI software development and assessment.

3. Analysis of the Variable Discrimination

Automated learning aims to mimic some of the natural learning processes existing in nature, the difference being that in automated learning, the learning is mainly based on a set of examples rather than following defined indications and rules that describe a given context. Similar to what happens with humans, ML often produces predictions and recommends decisions that end up being discriminatory to individuals or groups.

Among the available definitions of "Discrimination" in the context of ML and AI systems [9] is Verma and Rubin's approach describing discrimination as the direct or indirect relation between a protected attribute and the resulting prediction/classification/suggested decision. This is seconded by Mehrabi [10], where direct discrimination is distinguished by the direct relation between protected attributes and the produced prediction/classification/

decision with a negative consequence for the object being targeted by the decision. It expands by declaring that indirect discrimination not only relates to an indirect relation between the mentioned taxonomy but is also manifested when the implicit effects of protected attributes are considered. For instance, the use of an individual postal code in loan and insurance premium calculations are two examples showing how apparently less sensitive individual features may lead to a discriminatory decision.

According to Zhang, Wu, and Wu [11], residential areas often offer a representative distribution of their inhabitants in regard to attributes such as race, household income, etc. However, the zip code is not usually a protected attribute in the decision-making process because the law does not register it as a feature triggering discriminatory decisions like other features such as race or gender. In the literature, it is stated that a set of attributes the law suggests being treated as protected are exhibited in an attempt to help avoid discrimination in the aforementioned scenarios and others such as recruitment [12]. These examples allowed us to understand that discrimination is a variable that needs to be dealt with casuistically, in every new project, for every new and old scenario, across cultures.

It can be said that discrimination, in the context of ML and AI systems, has a statistical root when the information learned, by means of pattern discoveries, frequency measure, correlations among attributes, etc., about a group is used to judge an individual with similar characteristics. Hence, the importance of data and data collection procedures is carried out according to the scope of the intended decision or prediction.

The continued use of statistical methods in decision-making and/or the arrival of predictions leads to systematization of discrimination. Therefore, it can be understood that ML has scaled the impact of discrimination and "unintentionally institutionalized" these discriminatory methods through AI, and it has created a perpetual cycle where the object of discrimination itself becomes part of the knowledge base used in subsequent estimates, that, hence, become equally discriminatory. That is, a recommending software used within an enterprise with a given gender distribution will tend to reproduce the same unbalanced current gender distribution in their selection process while hiring new candidates. The referred distribution might not only be fit in correspondence to the enterprise's training base but also in correspondence with available knowledge about the top performers' distribution in the guild; the particular enterprise is part of what will result in perpetuating the gender distribution in the workforce and conditioning future hiring if the same method is used over time. This is the reason why discriminatory decisions are nowadays generally attributed to prediction, selection/estimation algorithms, etc. [13,14], and not to other equally important aspects such as data gathering, data cleaning, and data processing, for example.

Mehrabi et al. [10] add that discrimination can be classified as explainable and non-explainable according to the possibility of justifying or not justifying the produced decision/prediction from the triggering attributes. That is, explainable discrimination is close to what we understand as prejudice, where there is a clear parameter influencing the discriminatory decision or prediction. While non-explainable discrimination occurs when there is a discriminatory outcome that cannot be justified, the specific trigger cannot be identified. Either classification lacks ethical support.

Another study [15] conceptualizes discrimination in the context of ML and AI systems similarly to Mehari's study, and it justifies the use of these unintended discriminatory AI models by providing two main reasons: first, the model is able to provide a decision/prediction according to the need of the business; and second, the lack of a less discriminatory alternative model. This simply represents an attitude of resignation and acceptance of discrimination and the subsequent bias.

Additionally, in the literature [16], discrimination is defined using six classifications for bias:

1. Sample or selection bias, when the sample representativity becomes compromised with significant unbalance;
2. Measurement bias refers to systematic errors regarding data correctness, compromising the values supporting the estimations;

3. Self-reporting (survey) bias, related to the completeness of data, compromising the statistical significance and the accuracy backing the predictions;
4. Confirmation (observer) bias, resulting from the researcher's own prejudice while he or she presents information backing his or her working hypothesis;
5. Prejudice (human) bias, when the model/algorithm result reflects a pre-existent bias on the knowledge base used for training;
6. Algorithm bias, when the model/algorithm creates or amplifies bias from the training dataset in an attempt to overcome processing needs; this is usually true when working with multiple samples of different sizes.

As can be appreciated, discriminating upon the characteristics of an object is not intrinsic to humans. Technology reproduces and amplifies such behavior. The specialized literature exhibits a tendency to hold machine learning algorithms accountable for the problem created by their inability to adequately deal with bias, as analyzed in [3]; however, the data used in training and the data collection methods are equally responsible for discriminatory predictions and recommendations.

Lastly, it can be highlighted that discrimination has both an origin and cause of bias once the outcomes of today's discriminatory decisions based on yesterday's biases populate tomorrow's datasets. A visual aid can be found in Figure 1 below. In the field of the software industry, both variables, discrimination and bias, are closely related because of the speed at which the whole cycle occurs and because of the cycle's many iterations. The following section presents bias as a variable of analysis.

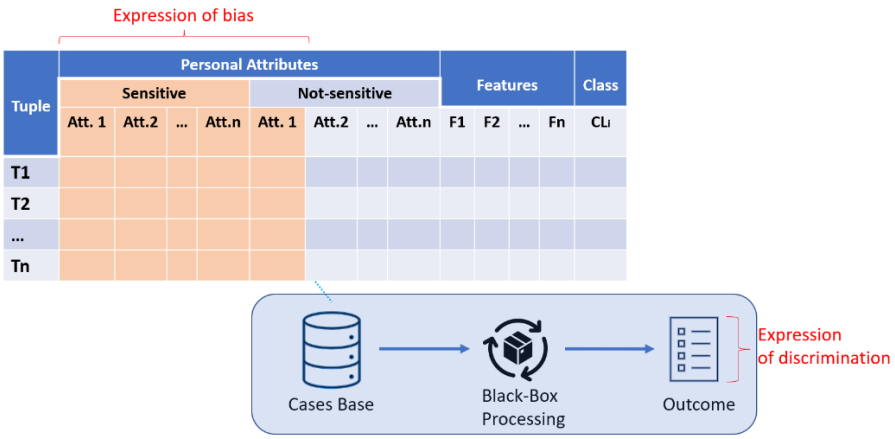

Figure 1. Simplified representation of an automated decision-making process.

4. Analysis of the Variable Bias

Similar to what occurs with human prejudice, the bias in ML leads to discriminatory predictions and recommendations. Consequently, many researchers are pursuing optimization of the methods in which ML identifies and eliminates bias. There are two marked methodological trends in that regard. The first trend pertains to algorithm calibration [17–24], while the most recent trends [25–30] aim to tackle the problem from the early stages of AI algorithms/model design.

Among the documents forming the Principled AI International Framework [31], the UNI Global Union 2017 report [32] describes bias as the action of using features such as gender, race, sexual orientation, and others as discriminatory elements in a decision with a negative impact somehow harmful to the human being. Then, the difference between bias concerning "Discrimination" is that "bias" represents the action while discrimination manifests itself in the result of using certain attributes in the decision-making process, as exhibited in Figure 1.

Figure 1 show how bias can be expressed in the inclusion of a subset of attributes oriented to the subject identification from the set of attributes describing a particular individual. Those attributes marked as an expression of bias in Figure 1 can be both

sensitive attributes, also referred to as protected (by researchers promoting the exclusion of such attributes from the decision), and insensitive attributes. The consideration of those attributes in the decision can result in a discriminatory outcome, as previously stated and also represented in the figure.

The dependence among these two variables could be located in this relation. It is also important to note that such a definition emphasizes the negative impact of the decision so that it seems not to consider "bias" when such an effect might be positive.

Another report, authored by the G20 [33], describes bias as the product of human activity with a given effect on individual rights and other contexts inherent to humans, while it declares that algorithms can unintentionally produce both bias and discrimination. The report also highlights the existence of two types of sources for bias: the method, either in the design of the algorithm or in the way the data is collected, and the distortion/corruption of the data used as the training basis for the model/algorithm.

We suggest the existence of two referents for the definition of bias within AIS: the statistical referent and the social referent. In that regard, Access Now Organization [34] presents the statistical referent as the distance between the AIS produced estimation/prediction and the actual occurrence of the estimated/predicted event. It explains that when there is statistical bias, there is evidence that the data represents a social bias, which is described as social bias by the same report.

Then, it is accurate to say that we are in the presence of an unfair dataset every time a discriminatory or biased conclusion is drawn and that any instance of an algorithm using that dataset for training will produce equally unfair decisions and predictions. That does not mean the same occurs in the opposite direction. The fact that an algorithm does not produce a discriminatory or biased decision/prediction does not indicate we are using a fair dataset. That, along with some related principles from the framework analyzed in [1,2], is the reason we suggest as a good practice (where applicable) the use of data pipeline dedicated frameworks to stress and exhaust datasets being used for algorithm and model training.

In that respect, the obligation of fairness defined by Access Now Organization [34] and The Public Voice Coalition [35] first suggests the existence of two benchmarks for the definition of bias in AI. The statistical reference is expressed as the deviation of the prediction in contrast with the event's actual occurrence, and the social reference is from the evidence of statistical bias within the data representing a social bias. Second, it recognizes that decisions/predictions reflecting bias and discrimination should not be normatively unfair. This means that decisions which are unfair and reflect biases must not only be assessed quantitatively but also evaluated with regard to their context with a case-by-case approach. This is to understand how to avoid them and create a norm/standard rather than being the exception to the rule. Additionally, third, it clarifies that the single evaluation of the outcomes (previously mentioned algorithm calibration) is not enough to determine the fairness of the algorithm or model. This idea was first explored in [3]. Consequently, Access Now Organization [34] and The Public Voice Coalition [35] propose the evaluation of pre-existing conditions in the data that can be further amplified by the AI system before its design is even considered. This report shows an inclination towards the emerging trend of recognizing in the data an origin for discriminatory and biased decisions, in contrast with the rooted trend of solely holding the algorithms accountable for the negative outcomes produced by AIS.

Additionally, the House of Lords Select Committee on Artificial Intelligence [36] and Martinho-Truswell et al. [37] criticize the methods of learning developed in machine learning, specifically how data is used during training. Per the House of Lords Select Committee on Artificial Intelligence [36], while learning, systems are designed to spot patterns, and if the training data is unrepresentative, then the resulting identified patterns will reflect those same patterns of prejudice and, consequently, they will produce unrepresentative or discriminatory decisions/predictions as well. Martinho-Truswell et al. [37] highlight that good-quality data is essential for the widespread implementation of AI technologies; how-

ever, the study argues that if the data is nonrepresentative, poorly structured, or incomplete, then there exists the potential for the AI to make the wrong decisions. Both reports define bias over the basis of misleading decisions produced from such compromised datasets.

Acknowledging the role of data in the introduction of bias is a relatively new approach (This is different from the Garbage In Garbage Out (GIGO) approach to explain the relation of trashy data input with faulty outputs. The GIGO approach links specific data issues such as duplicity of information, absence of information, and noise in information, just to provide a few examples, and bad programming with faulty output from systems. The relatively new approach of pointing out the datasets as an origin for discriminatory decisions refers to those datasets that, even when not being trashy, are biased and triggers discriminatory patterns in ADM systems. It is a new approach as the origin of discriminatory ADM systems' outcomes were mainly linked to biased algorithms, ignoring that datasets and the development team had a role in introducing bias into the system.). Mehrabi's [10] comprehensive survey provides several definitions of types of biases originating in the data. The author enriches upon the already mentioned historical and representation biases by providing further classifications. From the definitions provided by Mehrabi [10], we thought it pertinent to highlight the following due to the focus not on the data distribution per se but on the introduced bias resulting from a misuse of the dataset.

First, we wanted to note measurement bias, which takes place when using a particular feature of the object of the decision when building judgment, just because that feature has been historically over-measured. This particular action has a fuzzy line with human-introduced bias, as is explained later in the classifications provided by IBM.

The overall evidence shows that there exist some population groups that are more assessed and controlled (policed) than others and therefore have higher rates of arrests if we use the example of recidivism and risk assessment within the judicial system, turning those populations into groups vulnerable to this kind of bias.

Second, we wanted to point out the Evaluation bias that compromises the model validation when using inappropriate and disproportionated benchmarks in the verification process. The IJB-A benchmark known as the "Face Challenge" in face recognition was used to exemplify the matter because of its failures when considering skin color and gender.

There were four particularly interesting biases described in the study. First, aggregation bias, when false assumptions are made because of the use of conclusions produced by previously flawed models; the Simpson's Paradox related bias, referring to the different bias appreciations when looking at different data groupings within the analyzed dataset; the Linking bias, which arises when variables such as network sampling, method of interaction, and time are not considered when building a network around the object of the decision; and what they denominate, Emergent bias, resulting from user experiences with deployed products through the graphical user interface, where possible habits of prospective users were estimated from the design stages.

IBM [38] adds a human edge to the binomial data-algorithmic bias origin while presenting a set of unconscious bias definitions expressed in terms of their manifestation among the general population that engineers need to be consciously aware of when designing and developing for AI. Despite the IBM's classification in three main focus areas (shortcut biases, impartiality biases, and self-interest biases), we group those definitions into three main points of interest in project management as presented below. This new organization fits the context of our research as it moves the focus of IBM's classification from the individual to the project stage in which such biases can be introduced.

4.1. The First Point of Interest Is Project Conceptualization

We gathered IBM's Sunk Cost bias and Status Quo bias definitions under the project conceptualization point of interest. They both refer to the tendency to justify past choices and maintain the current situation, even though they no longer seem valid or when better alternatives exist. In that sense, AI practitioners need to be aware that every new project involves a unique business reality. Some highly specialized teams will try to accommodate

their expertise rather than study emerging methods when designing their solution approach. Sommerville [39] and the CHAOS report [40] stressed that issue is one of the main causes of project failure. Deciding the wrong project approach could be the first step toward an unfair AI system.

4.2. The Second Point of Interest Is Project Design

We gathered IBM's Not Invented Here bias, Self-Serving bias, and bias Blind Spot definitions under the design point of interest. We also divided this point of interest into two subcategories: Data affairs and Algorithm functioning affairs, as described below.

The Not Invented Here bias and the bias Blind Spot are somehow connected. The former refers to the aversion to contact with or use of products, research, standards, or knowledge developed outside the own group; and the latter refers to the tendency to see oneself as less biased than others or to be able to identify more cognitive biases in others than in oneself, something that might exhibit a cause-effect relation. The Self-Serving bias states the tendency to focus on strengths/achievements rather than on faults/failures. This suggests that AI practitioners should avoid discriminating against pre-existent approaches, which could save a significant amount of time and effort and provide valuable knowledge based not only on proven hypotheses but on errors or rejected hypotheses as well.

4.2.1. Data Affairs Subcategory

Under the Data affairs subcategory, we listed the Base Rate Fallacy, referring to the tendency to ignore general information and focus on specific information (a certain case), providing an individualistic opinion upon the decision's object. This is somehow related to the idea of stepping afar from generalizing based on previously available knowledge, given a group of subjects sharing some of their traits with the object of the decision. Additionally, the Availability bias focuses on overestimating events with greater "availability" in memory, influenced by how recent, unusual, or emotionally charged those memories might be.

4.2.2. Algorithm Functioning Affairs Subcategory

On the other hand, we listed the Congruence bias, Empathy Gap bias, Anchoring bias, and Bandwagon bias under the Algorithm functioning affairs subcategory.

The Congruence bias represents the tendency to test hypotheses exclusively through direct testing instead of testing alternative hypotheses. This approach ignores other variables that might affect the business being modeled, overlooking possible scenarios where the algorithm/model might behave differently regardless of the tested hypothesis's outcome.

Similarly, the Empathy Gap bias represents the tendency to underestimate the influence or strength of feelings in either oneself or others. This and the Congruence bias can be connected, whereas the inclination towards a given hypothesis ends up being accommodated.

Different from the Congruence and the Empathy Gap biases, the Anchoring bias relies almost entirely on one trait or piece of information when making decisions, usually the first piece of information that we acquire on the subject being targeted by the intended decision. It conceives a false illusion of objectivity when we separate ourselves from untested assumptions, such as our hypotheses and our feelings. However, the resulting decision ends up being biased because of the probable unrepresentativeness of the used data over the reality being modeled.

Finally, the Bandwagon bias portrays the tendency to do or believe things because many other people do. That kind of group thinking is wrong because following the general norm (when making decisions) contrary to making a decision as an individual might be forcing us to perpetuate bias. This is dangerous because doing so avoids the needed paradigm rupture in given situations, where the general historically agreed upon decisions are outdated.

4.3. The Third Point of Interest Is Project Verification and Validation

We then gathered Confirmation bias, Halo Effect, and Ingroup/Outgroup bias under the project verification and validation point of interest.

The Confirmation bias explains the tendency to search for, interpret, or focus on information in a way that confirms one's preconceptions. It might represent the previously referred connection between the empathy gap and congruence biases. Either way, the introduced bias, in this case, is supported by the under-representation of data used to reinforce one's own preconception.

The Halo Effect bias can be expressed by the predisposition of an overall impression to influence the observer. Positive feelings in one area cause ambiguous or neutral traits to be viewed adequately. This is not only important during the business modeling but also during verification tasks where the evaluator is too familiarized with the work being verified, measured, or audited.

The Ingroup/Outgroup bias describes the tendency or pattern to favor members of one's ingroup over outgroup members, favoring the institutionalization of bias.

Wrapping up the variable analysis, we can now state with support [41,42] that bias can be perceived as an intentional or unintentional predisposition toward prejudice in favor or against a person, object, or position. It has multiple origins within the context framed by the AI systems. Such origins include information represented within the data, the logic of algorithmic functioning, engineering methods and practices for data collection, data processing, and algorithmic design; it also can derive from intrinsic human biases for both designers and prospective users, and the contexts in which systems are used.

5. Analysis of Variable Fairness

By definition, heavy methodologies for software projects help developers and stakeholders to understand that efforts are needed along the software project lifecycle for verification and validation tasks. We can find several quality variables [39,43] that software projects have proactively managed in an attempt to avoid unintended outcomes from the systems they produce. Nowadays, with the use of AI systems, and particularly ML models and algorithms [44], consequential decisions are being automatically generated about people. The automation of bias, the incapacity of AI systems to bring neutrality to the decisions they produce, the perpetuation of bias, and the amplification of the historical discrimination are leading to concerns about how to ensure fairness. On one side, software practitioners strive to prevent intentional discrimination or failure, avoid unintended consequences, and generate the evidence needed to give stakeholders justified confidence that unintended failures are unlikely. On the other side, policymakers work to regulate the design and consumption of such systems so they are not harmful to human beings and that the necessary amendments are made in case they are required.

From a technical point of view, ref. [45] fairness is defined as the actions performed to optimize search engines or ranking services without altering or manipulating them for purposes unrelated to the users' interest. Expanding on that idea, in [32], it is acknowledged that fairness tasks should be planned during the design and maintenance phase of software development and that those tasks should seek to control negative or harmful human bias so that they are not propagated by the system.

Some studies [41,46] relate fairness to inclusion. For instance, ref. [41] stresses that fairness is expressed by means of inclusion and diversity by ensuring equal access through inclusive design and equal treatment. In [46], it is stated that AI systems should make the same recommendations for everyone with similar characteristics or qualifications. In consequence, software developers and software operators should be required to test the deployed solutions in the workplace on a regular basis to ensure that the system is built for its purpose and it is not harmfully influenced by bias of any kind—gender, race, sexual orientation, age, religion, income, family status, and so on—exposing the variable character of fairness over time. The report also states that AI solutions should adopt inclusive design efforts to anticipate any potential deployment issues that could unintentionally

exclude people. Both studies believe necessary the involvement of all affected stakeholders along the project lifecycle. This is a work philosophy that is shared by companies such as Telefónica [47], based in Spain, and one of the main telecommunication operators in Europe. Several of the techniques and metrics available describing how ML pursues fairness are mathematically formalized in the literature [9,10]. A critical analysis of metrics and techniques such as those formalized in both studies were criticized in [3].

A cultural attachment is also presented in [10] while defining the fairness variable when the authors state that different preferences and outlooks within different cultures condition the current situation of having multiple concepts for the term. The situation is aggravated by the fact that available definitions of fairness in philosophy, psychology, and computer science supporting algorithmic constraints are mostly based on Western culture. This led the authors to define fairness as the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making.

An even broader definition is being proposed by the Vatican [48] while using impartiality to explain fairness. The Vatican's working concept gathers the development and consumption of AI systems when it says, "*do not create or act according to bias*", and it connects the outcome of working to ensure fairness with its human focus when it says, "*safeguarding fairness and human dignity*".

To wrap up the analysis of the fairness variable, we wish to point out that these studies [9,10,32,46,48] define fairness as the AIS's ability to treat all similar individuals or groups equally and as the AIS's inability to produce harm in any possible way. This is indeed a noble but still very broad definition, and it shows the lack of agreement among the scientific community to achieve a definition of fairness that can be widely accepted. The Indian National Strategy for AI [49] locates the issue of fairness at the forefront of discussion in academic, research and policy fora, something that definitely merits a multidisciplinary dialogue and sustained research to come to an acceptable resolution, and it suggests identifying the in-built biases to assess their impact, and in turn, to find ways to reduce the biases until techniques to bring neutrality to data feeding AI solutions, or to build AI solutions that ensure neutrality despite inherent biases, are developed. In that regard, we need to stress that [10] indicates it is crucial to understand the different kinds of discrimination that may occur given the numerous distinct available definitions of fairness.

The analysis evidences a steering of the majority of the elements describing machine learning's traditional approach [17,20,23] to cope with bias and discrimination, moving away from its reactive character towards a more proactive style. Hence, it is appropriate to state that, in order to produce less discriminatory outcomes, in the context of AIS, the engineering focus needs to commute from fairness (as a nonfunctional requirement) to trustworthy AI as a business model.

6. Analysis of the Variable Trustworthiness

Several studies [4,6,7,50,51] agree that it requires human agency, oversight, and the use of a set of overlapping properties to define trustworthiness in the context of AI systems development and consumption. Among the most frequent highlighted properties across the studied bibliography, the following can be found:

1. Reliability is when the system does the right thing it was designed to and is available when it needs to be accessed.
2. Reproducibility is when the systems produce the same results in similar contexts.
3. Safety is when the system induces no harm to people as a result of their outcomes.
4. Security is when the systems are invulnerable or resilient to attacks.
5. Privacy is when the system protects a person's identity and the integrity of data, indicates access permission and methods, data retention periods, and how data will be destroyed at the end of such period, which ensures a person's right to be forgotten.
6. Accuracy is when the system performs as expected despite new unseen data compared to data on which it was trained and tested.

7. Robustness is when the system is sensitive to the outcome and to a change in the input.
8. Fairness is when the system's outcomes are unbiased.
9. Accountability is when there are well-defined responsibilities for the system's outcome such as the methods for auditing such outcomes.
10. Transparency is when it is clear to an external observer how the system's outcome was produced, and the decisions/predictions/classifications are traceable to the properties involved.
11. Explainability is when the decisions/predictions/classifications produced by the system can be justified with an explanation that is easy to be understood by humans while being also meaningful to the end-user.
12. Other variables such as data governance, diversity, societal and environmental well-being/friendliness, sustainability, social impact, and democracy.

Altogether, as supported by Brundage et al. [5], it can help build a trustworthy methodology to ensure users are able to verify the claims made about the level of privacy protection guaranteed by AI systems, regulators are able to trace the steps leading to a decision/prediction/classification and evaluate them against the context described by the modeled business, academics are able to research the impacts associated with large-scale AI systems, and developers are able to verify best practices are set for each of the AI development stage within the project lifecycle.

In order to achieve Trustworthy AI, the Independent High-Level Expert Group on AI [41] recommends enabling inclusion and diversity throughout the entire AI system's development project's life cycle involving all affected stakeholders throughout the process. Along with Abolfazlian [50], both studies describe three components trustworthy AI should comply with throughout the system's entire life cycle: it should be lawful, complying with all applicable laws and regulations; it should be ethical, ensuring adherence to ethical principles and values; and it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm. Similarly, Gagnon [4] proposes three other main components trustworthy AI systems should consist of the following:

- Ethics of algorithms (respect for human autonomy, prevention of harm, fairness, explicability);
- Ethics of data (human-centered, individual data control, transparency, accountability, equality), and;
- Ethics of Practice (responsibility, liability, codes, regulations).

This actually represents an attempt to harness unintended discrimination produced by AIS, from the perspective of the policymaking and legal norms, specifically with a basis on the International Law of Human Rights. Given that engineering methods alone could not be sufficient enough to protect, according to Fjeld et al. [31], the fundamental rights from unintended harms of AI systems. As seen above, the Principled AI International Framework presented by Fjeld et al. [31] gathers a global effort to establish a set of policies and guidelines informed by principles as a methodological reference when designing AI. Despite the progress that this mechanism might represent from the legal point of view, it is yet insufficient as a methodological mechanism manageable by AI designers given their background and the language [28,29] discrepancies among legal jargon and the software profession, better detailed in [1,2].

7. Threats to Validity

The scope of the exploration performed in the present study on variables such as bias, discrimination, fairness, trustworthiness, and others, is aligned with the working definitions from Intergovernmental organizations (G20, UNI Global Union, European Union, etc.), Governments (USA, France, UK, India, UAE, etc.), the private sector (Telefónica, Microsoft), the public sector (Universities, Access Now Org, The Public Vice Coalition, etc.), Religious authorities (Vatican), and other referenced researchers, within a given timeframe, that recognize the ethical and social dimensions of such variables along to their usual functional extent. The authors acknowledge the existence of multiple definitions beyond the analysis

presented in this paper and wish to assure that the current and future existence of those other definitions does not threaten the validity of our work but rather enriches it. In that same line of thought, the authors find it pertinent to highlight the direct relation of the studied variables with societies 'shared values, and with time, as the main reasons to support the emergence of modified views and new assertions for the variables.

8. Conclusions

This study shows the lack of agreement among the scientific community in reaching a standardization of the studied variables to support trustworthy AI as a business model to be assimilated by software developers, especially by AIS designers, when designing AIS. That could be other of the reasons, along with the ones flagged already for the Principled AI International Framework principle's ambiguities described in previous studies.

Discrimination and bias are two entangled variables with a strong interdependency that results in one of them being the cause and the effect of the other. For the purposes of the present study, bias refers to the action of deciding upon an individual or group with a given potentially harmful impact because of their features, while discrimination is expressed by the outcome of the decision itself.

Discrimination, and by extension, fairness, are culture dependable variables. In that regard, there must be required a dedicated assessment for every new project during the conceptualizing stage regardless of the scenario and how the variables will behave across cultures in which the projected ADM system will be deployed.

The study shows that ADMS's biased and discriminatory outcomes are not only a consequence of faulty algorithms and models but are also linked to other processes such as data gathering, data cleaning, and data processing; and also conditioned by the development team's own bias.

The study also identifies the main variables that principled AI suggests trustworthy AI should be built upon (through fairness and non-discrimination). Consequently, references for methodological approaches to the implementation of trustworthiness, in the context of ADMS, could be orchestrated through fairness and non-discrimination as specific goals and based on the following four derived features: (1) transparency, which involves specifically related variables such as explainability and accountability; (2) security, that involves specifically related variables such as safety and privacy; (3) project governance, involving specifically related variables such as environmental commitment, societal wellbeing, diversity and inclusion, sustainability, social impact, and compliance with the law and regulatory norms; and (4) bias management, that involves specifically related variables such as knowledge transfer, training, and data collection. These features and their derived related variables are taken as checkpoints when determining the level of capability and maturity the AIS designers achieve in their development process when building a trustworthy product or system.

The variables explored in this study allow us to define the environment in which a methodological reference tool for the development of trustworthy ADMS should suit. Therefore, the analysis of these variables, and their relations, described in the previous conclusion, provides the basis for our future lines of work. That is the design of the main, general, and specific goals, as well as the quality features for a proposal of a Capability and Maturity Model for the development of trustworthy ADMS. With a Capability and Maturity Model for the development of trustworthy ADMS, we aim to support software engineers, specially ADMS developers, to incorporate into their current working definitions the social and ethical dimensions of those variables.

Author Contributions: Conceptualization, D.V. and J.L.S.; methodology, D.V.; investigation, D.V.; writing—original draft preparation, D.V.; writing—review and editing, J.L.S.; supervision, J.L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Varona, D.; Suarez, J.L. Analysis of the principled-AI framework's constraints in becoming a methodological reference for trustworthy-AI design. In *Handbook of Computational Social Science*; Engel, U., Quan-Haase, A., Xun Liu, S., Lyberg, L.E., Eds.; Routledge Taylor and Francis Group: Oxfordshire, UK, 2022; Volume 1, ISBN 9780367456528.
2. Varona, D.; Suarez, J.L. Principled AI Engineering Challenges Towards Trust-worthy AI. *Ethics Inf. Technol.* 2022, *submitted*.
3. Varona, D.; Lizama-Mue, Y.; Suarez, J.L. Machine learning's limitations in avoiding automation of bias. *AI Soc.* **2020**, *36*, 197–203. [CrossRef]
4. Gagnon, G.P.; Henri, V.; Fasken; Gupta, A. Trust me!: How to use trust-by-design to build resilient tech in times of crisis. *WJCOMPI* **2020**, *38*, 1–6.
5. Brundage, M.; Avin, S.; Wang, J.; Belfield, H.; Krueger, G.; Hadfield, G.K.; Khlaaf, H.; Yang, J.; Toner, H.; Fong, R.; et al. Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv* **2020**, arXiv:2004.07213.
6. Wickramasinghe, C.S.; Marino, D.L.; Grandio, J.; Manic, M. Trustworthy AI Development Guidelines for Human System Interaction. In Proceedings of the 13th International Conference on Human System Interaction, Tokyo, Japan, 6–8 June 2020.
7. Smith, C.J. Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development. In Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
8. Mittelstadt, B. Principles alone cannot guarantee ethical AI. *Nat. Mach. Intell.* **2019**, *1*, 501–507. [CrossRef]
9. Verma, S.; Rubin, J. Fairness Definitions Explained. In Proceedings of the International Workshop on Software Fairness (FairWare 2018), Gothenburg, Sweden, 29 May 2018.
10. Mehrabi, N.; Morstatter, F.; Saxena, N.A.; Lerman, K.; Galstyan, A.G. A survey on bias and fairness in machine learning. *Mach. Learn.* **2019**, *54*, 1–35. [CrossRef]
11. Zhang, L.; Wu, Y.; Wu, X. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017.
12. Chen, J.; Kallus, N.; Mao, X.; Svacha, G.; Udell, M. Fairness Under Unawareness: Assessing Disparity when Protected Class is Unobserved. In Proceedings of the Conference on Fairness, Accountability and Transparency, Atlanta, GA, USA, 29–31 January 2019.
13. Jago, A.S.; Laurin, K. Assumptions about algorithms' capacity for discrimination. *Personal. Soc. Psychol. Bull.* **2021**, 1–14. [CrossRef] [PubMed]
14. Loi, M.; Christen, M. Choosing how to discriminate: Navigating ethical trade-offs in fair algorithmic design for the insurance sector. *Philos. Technol.* **2021**, *34*, 967–992. [CrossRef]
15. Schmidt, N.; Siskin, B.; Mansur, S. How Data Scientists Help Regulators and Banks Ensure Fairness when Implementing Machine Learning and Artificial Intelligence Models. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 19 June 2018.
16. Martínez-Plumed, F.; Ferri, C.; Nieve, D. Fairness and missing values. *arXiv* **2019**, arXiv:1905.12728.
17. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* **2017**, *5*, 153–163. [CrossRef] [PubMed]
18. Feldman, M.; Friedler, S.A.; Moeller, J.; Scheidegger, C.E.; Venkatasubramanian, S. Certifying and Removing Disparate Impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sidney, Australia, 10–13 August 2015.
19. Fish, B.; Kun, J.; Lelkes, Á.D. A Confidence-Based Approach for Balancing Fairness and Accuracy. In Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, FL, USA, 5–7 May 2016.
20. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3315–3323.
21. Reich, C.L.; Vijaykumar, S. A Possibility in Algorithmic Fairness: Calibrated Scores for Fair Classifications. In Proceedings of the 2nd Symposium on Foundations of Responsible Computing (FORC 2021), Virtual Event, 9–11 June 2021. arXivLabs:2002.07676.
22. Pedreschi, D.; Ruggieri, S.; Franco, T. *Discrimination-Aware Data Mining Technical Report: TR-07-19*; Dipartimento di Informatica, Universitµa di Pisa: Pisa, Italy, 2007.
23. Solon, B.; Selbst, A.D. Big data's disparate impact. *Calif. L. Rev.* **2016**, *104*, 671–732.
24. Zafar, M.B.; Valera, I.; Rodriguez, M.G.; Gummadi, K.P. Fairness Constraints: Mechanisms for Fair Classification. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), San Diego, CA, USA, 9–12 May 2015. arXiv:1507.05259.
25. Holstein, K.; Vaughan, J.W.; Daumé, H.; Dudík, M.; Wallach, H.M. Improving Fairness in Machine Learning Systems: What do Industry Practitioners Need? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019.
26. Varona, D. La responsabilidad ética del diseñador de sistemas en inteligencia artificial. *Rev. Occidente* **2018**, *446–447*, 104–114.
27. Varona, D. AI systems are not racists just because. In Proceedings of the T-13 hours: Building Community Online in CSDH/SCHN2020, Virtual Event, 1–5 June 2020.

28. Varona, D. (Western University, Canada). Artificial Intelligence Design Guiding Principles: Review of "European Ethical Charter on the Use of AI in Judicial Systems and Their Environment". 2020. Available online: https://www.danielvarona.ca/2020/06/17/artificial-intelligence-design-guiding-principles-review-of-european-ethical-charter-on-the-use-of-ai-in-judicial-systems-and-their-environment/ (accessed on 1 July 2021).
29. Varona, D. (Western University, Canada). Artificial Intelligence Design Guiding Principles: Review of "Recommendation of the Council on Artificial Intelligence". 2020. Available online: https://www.danielvarona.ca/2020/06/28/artificial-intelligence-design-guiding-principles-review-of-recommendation-of-the-council-on-artificial-intelligence/ (accessed on 1 July 2021).
30. Veale, M.; Van Kleek, M.; Binns, R. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–27 April 2018.
31. Fjeld, J.; Achten, N.; Hilligoss, H.; Nagy, A.; Srikumar, M. *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*; Berkman Klein Center for Internet & Society: Cambridge, MA, USA, 2020.
32. UNI Global Union. *The Future World of Work. Top 10 Principles for Ethical Artificial Intelligence*; UNI Global Union: Nyon, Switzerland, 2017.
33. Abrieu, R.; Aneja, U.; Chetty, K.; Rapetti, M.; Uhlig, A. The Future of Work and Education for the Digital Age: Technological Innovation and the Future of Work: A View from the South. Technical Report. Argentina: G20. July 2018. Available online: https://www.g20-insights.org/wp-content/uploads/2018/07/GSx-TF-1-PB-Albrieu-et-al-final-2.pdf (accessed on 19 May 2021).
34. Access Now Organization. Human Rights in the Age of AI. Technical Report. AccessNowOrg. November 2018. Available online: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf (accessed on 19 May 2021).
35. The Public Voice Coalition. *Universal Guidelines for AI*; The Public Voice Coalition: Geneva, Switzerland, 2018.
36. House of Lords Select Committee on Artificial Intelligence. AI in the UK: Ready, Willing and Able? Technical Report. Authority of the House of Lords. 2018. Available online: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf (accessed on 19 May 2021).
37. Martinho-Truswell, E.; Miller, H.; Nti Asare, I.; Petheram, A.; Stirling, R.; Gomez Mont, C.; Martinez, C. Towards an AI strategy in Mexico: Harnessing the AI revolution. *White Pap.* **2018**, 23.
38. IBM. *Everyday Ethics for AI*; IBM: Armonk, NY, USA, 2019.
39. Sommerville, I. *Software Engineering*; Pearson Education: London, UK, 2016.
40. The Standish Group. *CHAOS Report 2020*; The Standish Group: Boston, MA, USA, 2020.
41. Independent High Level Expert Group on AI. *AI Ethics Guidelines for Trustworthy AI*; European Commission: Brussels, Belgium, 2019.
42. Smart Dubai Office. *AI Ethics, Principles and Guidelines*; Smart Dubai Office: Dubai, United Arab Emirates, 2019.
43. Pressman, R. *Software Engineering. A Practitioner's Approach*, 7th ed.; McGrawHill Higher Education: New York, NY, USA, 2010; p. 930.
44. National Science and Technology Council, Committee on Technology. *Preparing for the Future of Artificial Intelligence*; Executive Office of the President: Washington, DC, USA, 2016.
45. Demiaux, V.; Si, A.Y. *How Can Humans Keep the Upper Hand? The Ethical Matters Raised by Algorithms and Artificial Intelligence*; The French Data Protection Authority (CNIL): Paris, France, 2017.
46. Task Force 7. The Future of Work and Education for the Digital Age. Technical Report. T20. 2020. Available online: https://t20japan.org/task-forces/the-future-of-work-and-education-for-the-digital-age/ (accessed on 19 May 2021).
47. Telefónica. *AI Principles of Telefónica*; Telefónica: Madrid, Spain, 2018.
48. Servizio Internet Vaticano. *Rome Call for AI Ethics*; Servizio Internet Vaticano: Vatican City, Vatican, 2020.
49. NITI Aayog. *National Strategy for Artificial Intelligence*; NITI Aayog: New Delhi, India, 2018.
50. Abolfazlian, K. Trustworthy AI needs unbiased dictators! *Artif. Intell. Appl. Innov.* **2020**, *584*, 15–23. [CrossRef]
51. Wing, J.M. Trustworthy AI. *arXiv* **2020**, arXiv:2002.06276. [CrossRef]