Fairness and Bias in Learning Systems: a Generative Perspective

Serge Dolgikh

National Aviation University, 1 Lubomyra Huzara Ave, Kyiv, 03058, Ukraine

Abstract

In this work that is in progress we approach definitions and analysis of fairness and bias in the learning systems from the perspective of unsupervised generative learning. Based on generative structure of informative low-dimensional representations that can be obtained, as demonstrated previously, with different types and architectures of unsupervised generative models, certain types of bias analysis can be performed without massive prior (True Standard) data. As demonstrated on examples, these methods can provide additional angles and valuable insights in the analysis of bias and fairness of learning systems.

Keywords 1

Learning systems, bias, unsupervised learning, generative learning, clustering

1. Introduction

Whereas AI technology has been developing at an outstanding pace, finding applications in many areas and domains of research, industry and societal functions, the progress has not been entirely and unconditionally positive. One direction of questioning is understanding the reasoning of AI systems and ability to provide explanations or audit of their decisions ("black box" vs. explainable AI, [1,2]). Another, though closely related one is developing conceptual and ontological framework describing the fairness, trustworthiness and bias of AI systems.

Many studies described examples of bias in different functional applications of AI, including criminal justice, health care, human resources, social networks and others [3,4]. It was noted that the problems of explainable and trusted AI are closely interrelated: understanding the reasons why learned systems make certain decisions can be a key factor in determining whether they can be trusted; on the other hand, it is not easy to imagine a mechanism or a process of confident and reliable determination of a trusted AI without some insight into the reasons of its decisions.

In this work we pursue a perspective on these essential and actual questions that does not involve prior trusted data for such determination. This approach allows to unbind the question of "chicken and egg" or bootstrap in determination of trustworthiness (the origins of the trusted system that produced prior decisions) while suggesting sound and practical methods of analysis of fairness and bias based on generative structure of the input data. In our view, methods of unsupervised generative concept learning that are being actively developed [5] can provide a basis for such analysis.

In essence, methods of generative concept learning, where successful, can establish a structure of characteristic patterns, types or natural concepts in the input distribution by stimulating learning models to improve quality of generation from informative latent distributions, often of significantly reduced complexity. Unlike methods of conventional supervised learning, these approaches do not depend on specific assumptions about the distribution of the data, massive sets of prior data and can be used in a general process with data of different types, origin and domains of application.

Assuming that generative structure of the data of interest has been obtained, an analysis of distributions of decisions produced by the audited AI systems across characteristic natural classes of input data can provide valuable insights about the system, including possibility of bias.

1st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-22, co-located with AIxIA 2022, University of Udine, Udine, Italy, 2022
EMAIL: sdolgikh@nau.edu.ua (A. 1)
ORCID: 0000-0001-5929-8954

2. Bias and Trustworthiness: Approaches and Definitions

We will consider the black box interpretation of AI (and generally, a learning system, LS) whereby a functional or trained LS L can produce decisions on the set of inputs S, for example, sensory inputs from the environment:

$$d(x) = D(x, L), \qquad x \in S \tag{1}$$

where $D(x)$, the decision function of the learning system, however, the justification or explanation for a particular decision $d(x)$ is not necessarily known to the external observer.

On a subset of inputs, presumably representative sample of the input distribution, the system produces a set of decisions, $D(S)$.

In one approach, suppose there exists a True Standard ("etalon", standard, TS) set of outputs that represents correct decisions for given inputs with sufficient confidence. Then, characteristics of the trained system such as accuracy and error can be defined with standard measures based on the TS decisions by comparing the decisions produced by the system with those in the standard set.

$$A, E(L): \{ D(S), TS \} \tag{2}$$

Where A, $E(L)$: accuracy and error of the learning system over the representative set of inputs and presumably, the input space S.

Definition of the bias on the other hand, is not as straightforward. To begin, some observations though trivial need to be made.

1. Bias can be defined on a system level, for example, a subset of decisions $D(S_X, L)$ and not on an individual decision; same decision on the same input can be produced by an unbiased and biased system.
2. It can be argued that with a black box system, biased correct decisions are indistinguishable from unbiased ones. For example, it is common to see trained AI system biased to acceptance or rejection; such a bias can be easily detected with an adequate etalon set. However, if one considers only the subset of correct decisions, no conclusion about the bias of the system can be made. As a consequence of this observation, bias analysis in the case of black box learning systems, where additional context of decisions (explanations) is not available, the analysis has to be limited to the subset of incorrect decisions: $S_w: D(x) \neq TS(x), x \in S_w$.
3. Next, bias is not equivalent to wrong decisions, errors. As already mentioned, there is no reason to expect that correct decisions cannot be produced by biased systems (i.e., correct decisions made for "wrong reasons"); also, unbiased systems can produce incorrect decisions (errors).

Based on these observations, a bias in a learning system can be defined as a systematic deviation of decisions produced by the system from correct (etalon) decisions correlated with a set of certain factors (bias factors). The relationship between accuracy and trustworthiness are illustrated in **Figure 1**.

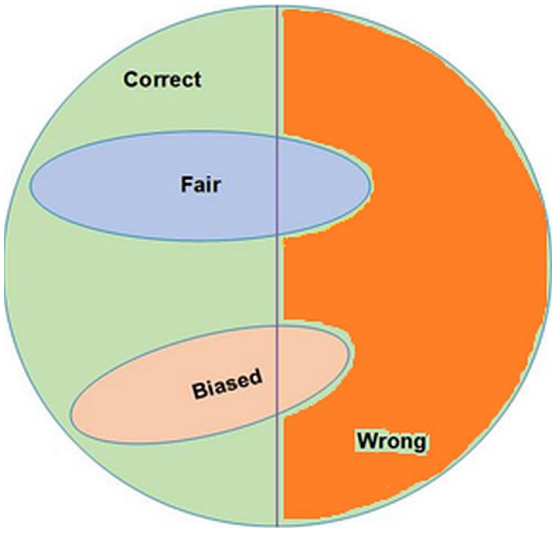

Figure 1: Bias, Fairness and Accuracy.

Errors include bias but not limited to it (imperfections, failure to learn). As well, biased systems can produce correct decisions. Consequently, determination of bias can be more challenging than that of correctness that can be measured by standard metrics of accuracy at least in the domain of learning with known TS decisions such as conventional supervised learning.

2.1. Challenges and Approaches in Measuring Bias and Trustworthiness

In approaching the question, how trustworthiness and bias can be measured, evaluated for realistic learning systems these challenges can be encountered:

1. True Standard (TS) decisions may not be available for all or significant part of the inputs or be insufficient to make confident judgements.
2. Can TS decisions themselves be trusted? (i.e., assured to be free of bias)?

The second question can be far from trivial as has been discussed in numerous studies. These questions relate to "conventional" approach in determining the bias based on prior trusted True Standard decisions. However, it may not be the case in all cases and domains, that brings another perspective:

3. Is definition of bias possible that *is not based on pre known TS decisions*?

This question parallels the dichotomy of supervised versus unsupervised learning: where successful learning can be dependent on prior sets of successful decisions (conceptual bootstrap problem).

Thus, in exploration of bias in learning systems one can outline two broad directions:

1. Analysis of fairness and bias based on available True Standard decisions.

2. Approaches in evaluation of bias / trustworthiness without resorting to TS decisions that may not be available for specific task or problem area.

In the rest of this work we will focus our attention on the second problem area.

2.2. Non-Standard Bias Analysis

In scenarios where standard decisions for evaluation of bias and trustworthiness are not available alternative approaches need to be developed. Clearly, evaluation of correctness of LS decisions is not possible without some measure, or criteria that are imposed externally. Same decision can be correct or the opposite, if different sets of criteria are applied. This, the first essential input to these methods is the correctness criteria.

Secondly, it will be assumed that the data used to created (for example, train) the learning system is a correct representation of its sensory environment. Of course, it does not guarantee that the environment itself is correct, that is, representative and fair representation of some desired purpose or objective; such scenarios fall beyond the scope of the study.

Based on these assumptions, the problem of detection of bias that is not based on availability of TS (non-standard bias analysis) can be formulated as: *determine the probability of systemic deviation of system decisions from the input criteria, correlated with one or several bias factors*.

For a representative subset of decisions D of a functional (trained) LS L on the subset of inputs S, and a small set of criteria C, determine the probability $p_B(L)$ of L being biased; secondly, attempt to identify the bias factors $f_B(L)$ correlated with the biased decisions $D_B \subset D$.

$$D, C \rightarrow p_B(L), f_B(L) \tag{2}$$

"small" here could mean that cardinality of the criteria set has to be much lower than that of a reasonable set of standard decisions: $Card(C) \ll Card(TS)$.

In the remaining sections we will attempt to illustrate how this program can be realized based on the ability of certain learning systems to create informative *generative* representations of the sensory data that do not require massive prior sets of standard decisions to derive certain essential information from observed distributions. A distinct feature of such system is an ability to learn from the incentive to improve perceptions, or generations of observable inputs in a process of self-supervised learning [6] that, as a number of results have demonstrated, can lead to emergence of characteristic conceptual structure in the resulting representations of sensory data [7,8].

2.3. Generative Representations and Non-Standard Bias Analysis

As has been reported in a number of results, models of unsupervised generative learning can produce informative representations of complex sensory data of different types and origin with clear conceptual structure and significant reduction of dimensionality [7,8]. An example of two-dimensional generative representations of a dataset of images of basic geometric shapes is given in **Figure 2**.

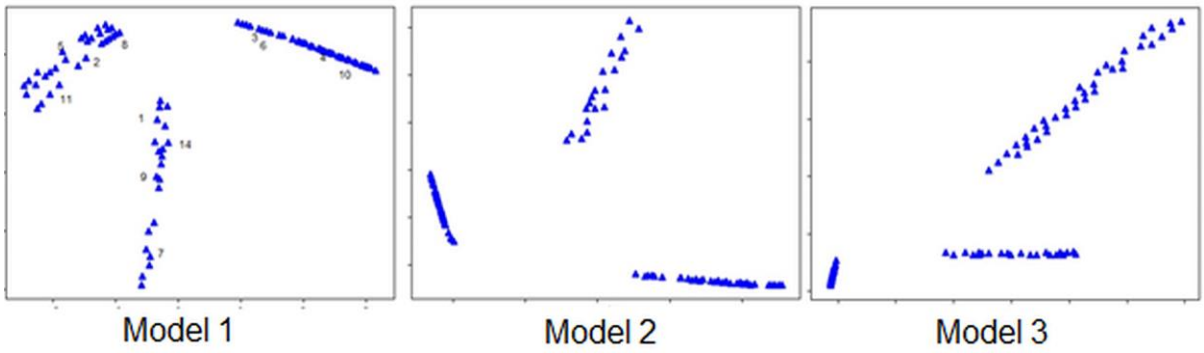

Figure 2: Conceptual representations, set of geometrical shapes.

In the illustration, two-dimensional latent representations of a set of images of basic geometric shapes: circles, triangles and empty backgrounds were plotted in the latent coordinates with three independently learned models of unsupervised generative learning [9]. Though without prior knowledge one cannot make any conclusions about the semantics of the input data from which representations were obtained, it is clear from the distribution of the encoded data that it contained at least three distinct types, patterns or concepts. Essentially, successful generative learning allows to identify characteristic structure of arbitrary data by factorizing the latent distribution into characteristic regions (natural concepts). An example of such factorization can be observed in the figure above.

An essential advantage of these methods is that they are entirely unsupervised, that is, do not require any prior knowledge of the data, and as such would be in compliance with the objective of non-standard bias analysis as defined earlier. Such unsupervised decomposition of data into characteristic latent structures, if and where successful, can provide an additional perspective for bias analysis, allowing to bypass the dependency on massive standard decision sets.

Indeed let us consider a structure of latent clusters $K_S = \{ K_n \}$ in the representative input set S identified with certain level of confidence, γ and decisions produced on it by a black box learning model: $D(S)$. With the latent structure K_S the decision set can be decomposed into distributions over the identified clusters as: $D_S = \{ D(x), x \in K_n \}$. In contrast to decomposition by observable parameters that in a real complex data can be of very high dimensionality, the advantages of generative decomposition are: 1) significantly lower dimensionality of the latent generative space and 2) generative factorization that represents characteristic, or natural types, classes or patterns in the input data.

A comparison of distributions of decisions D_S across different natural clusters can provide additional and independent perspective for an analysis of possible bias.

For an illustration, let us consider the set of geometrical shapes above, presuming that it describes some observable data on which decisions are produced by a black box learned system L. As the inputs to bias analysis, one would have the set of decisions D_S produced by L on a representative set of inputs S, in some format, suppose for simplicity, Boolean or real number representing probability.

A common approach in conventional bias analysis would be to seek a correlation between the decisions and observable parameters, and examination of such correlations for potential bias. A large number of observable parameters (i.e., dimensionality of the data samples in the set) can present significant challenges with such an approach, as well as a possibility of a more complex correlation with multiple input parameters that may not be easily detected.

To illustrate application of generative methods in this example, suppose unsupervised generative models denoted produced a consistent decomposition of the dataset into characteristic clusters K_S with the distribution of decisions in the identified clusters, D_S: $D = (K_s, D_s)$. An example of such distribution

for two learning systems denoted “F” and “B”, of similar overall accuracy, is shown in Table 1. The level of trustworthiness or bias of each system is not known at this stage in the bias analysis.

Table 1

Generative decomposition of decisions, shapes dataset

System	Mean decision			
	Concept 1	Concept 2	Concept 3	Dataset
“F”	0.18	0.28	0.16	0.18
“B”	0.28	0.17	0.11	0.20

Once distributions of decisions by characteristic clusters in the input data is obtained, they can be examined for possible bias. Of many possibilities, we will outline two.

In one case, suppose significant differences are observed between distributions of decisions in the clusters, as illustrated in Table 1 (*inter-concept decision disparity*). As discussed earlier, the test of fairness depends on defined criteria of correctness and let us suppose in this case the hypothesis or objective for the test of fairness is defined as: “*no significant differences in decisions observed between identifiable groups of subjects*”. From the results of generative bias analysis above, obtained without any TS decisions, one can observe that such differences can be seen in both models: Model “F”: Concept (Cluster) 2; Model “B”: Concepts 1, 3 and additional analysis is necessary.

Next, one can examine representative samples of clusters that also can be obtained from generative analysis (**Figure 2**) and investigate whether deviation from average decision is “justified”, that is, can be explained for these samples based on the objective. Suppose the additional analysis produced this outcome:

Concept 1: “*No*” (disparity of the group from the set mean is not justified)

Concept 2: “*Yes*” (deviation from the mean is justified or explainable)

Concept 3: “*No*”

Based on this analysis, one can conclude that model “F” satisfied the generative test of fairness whereas system “B” failed it (by producing decisions incompatible with the objective). Moreover, the factors of bias can be identified in this case as latent coordinates of the regions of diverging clusters.

In another case, suppose a learning system have developed a spurious bias with one or some of the input parameters (so called “training shortcut”). This would cause presence of outlier points with expressed deviations from cluster means in some clusters, with observable parameters associated with the bias condition (*intra-concept decision disparity*). Then the outlier set can be examined for justified deviation as in the preceding case, resulting in confirmation or discarding of the bias hypothesis. Again, correlation analysis of the observable parameters with the outlier samples may indicate the bias factors that were developed in the training process.

3. Conclusions

As discussed in this work, unsupervised generative analysis of observable data and the structure of natural types or concepts it can produce, can provide additional perspective and inputs for the analysis of bias and fairness of black box learning systems. An analysis based on a structure of natural types that can be identified with entirely unsupervised methods can bypass the requirement for massive prior True Standard decisions common with conventional methods of machine intelligence, while providing a basis for confident determination of possible bias in the discussed scenarios of inter- and intra-cluster disparity of the decisions. Due to high versatility of models and architectures of generative learning, including deep neural networks, the method can have a broad range of applicability in problems and with data of different types.

It is important to remember however that it is only an additional approach in analysis of possible bias that does not and cannot make a claim to a final determination. Generally, at least in the defined context, it can be challenging to guarantee an absence of bias as it would be equivalent to a negative proof of absence of correlation of an arbitrary set of decisions with *any* factor. Nevertheless, unsupervised generative analysis can offer some valuable insights in this increasingly actual domain of applications of Artificial Intelligence.

4. References

[1] Longo L., Goebel R., Lecue F., Kieseberg P., Holzinger A.: Explainable Artificial Intelligence: concepts, applications, research, challenges and visions. CD-MAKE 2020, LNCS 12279, 1–16, (2020).

[2] Schwartz R., Vassilev, Green K., Perine L., Bart A.: Towards a standard for identifying and managing bias in Artificial Intelligence. National Institute of Standards and Technology, USA Special Publication 1270 https://doi.org/10.6028/NIST.SP.1270 (2022).

[3] Bogen, M.: All the ways hiring algorithms can introduce bias. Harvard Business Review (2019).

[4] Gianfrancesco M.A, Tamang S., Yazdany J., Schmajuk G.: Potential biases in Machine Learning algorithms using electronic health record data. JAMA International Medicine, 178(11), 1544 (2018).

[5] Bengio, Y.: Learning deep architectures for AI. Foundations and Trends in Machine Learning 2(1), 1–127 (2009).

[6] Jing L., Tian Y.: Self-supervised visual feature learning with deep neural networks: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (11), 4037-4058 (2021).

[7] Higgins, I., Matthey, L., Glorot, X., Pal, A., et al.: Early visual concept learning with unsupervised deep learning. arXiv1606.05579 [cs.LG] (2016).

[8] Dolgikh, S.: Low-dimensional representations in generative self-learning models. In: Proc. 20th International Conference Information Technologies – Applications and Theory (ITAT-2020), Slovakia, CEUR-WS.org 2718, 239–245 (2020).

[9] Dolgikh, S.: Topology of conceptual representations in unsupervised generative models. In: Proc. 26th International Conference Information Society and University Studies (IVUS-2021) Kaunas Lithuania, CEUR-WS.org 2915, 150–157 (2021).