Is diverse and inclusive AI trapped in the gap between reality and algorithmizability?

Carina Geldhauser*[1] and Hermann Diebel-Fischer[2]

[1]Munich Centre for Machine Learning and Technische Universität München, Boltzmannstr. 3, 85748 Garching
[2]ScaDS.AI and Technische Universität Dresden, 01062 Dresden
carina.geldhauser@ma.tum.de, hermann.diebel-fischer@tu-dresden.de

Abstract

We investigate the preconditions of an operationalization of ethics on the example algorithmization, i.e. the mathematical implementation, of the concepts of fairness and diversity in AI. From a non-technical point of view in ethics, this implementation entails two major drawbacks, (1) as it narrows down big concepts to a single model that is deemed manageable, and (2) as it hides unsolved problems of humanity in a system that could be mistaken as the 'solution' to these problems. We encourage extra caution when dealing with such issues and vote for human oversight.

1 Introduction

In the past years, a high number of AI ethics guidelines have been published, see e.g. the reviews [1, 2]. In addition, a myriad of publications, from books, white papers, policy briefs to blog posts, has appeared, some of those with the promise to lead us to the holy grail of the perfect, flawless, and yet ethical-by-human-standards AI system.

Soon after, these guidelines and principles were criticized as 'ineffective', 'meaningless' or 'toothless' [3, 4], which often meant that a clear operationalization of the high-level principles described was missing.

Ideas or guidelines for potential operationalization, organizational awareness and the call for more ethics education of developers were published in high numbers as well, e.g. [5, 6], but the gap between algorithmically efficient and feasible models and the complex reality remained. To give an example, a recent article by economists concluded

> there is an apparent gap between the results of numerous tools and the formal requirements to deem a risk sufficiently mitigated or controlled. This gap between tools and abstract trustworthiness requirements should be addressed by future research.[1]

There are indeed many questions left open by current operationalization proposals, which need to cater diverse stakeholders. While pragmatic voices from industry call to explore trade-offs between accuracy, efficiency, fairness or diversity[2] there is a general disagreement on whether a fundamental ethical problem can be "audited away". Research also revealed that

> well intentioned attempts at algorithmic auditing can have effects that may harm the very populations these measures are meant to protect.[3]

In summary, consensus is missing not only the operationalization of "ethical AI", but also the definition of trustworthiness or fairness. Arguably, a business-oriented definition is simply "avoiding unwanted side-effects" [9].

In this article, we argue that trustworthiness in AI is substantially more: it involves a human factor. This human factor results from the blurriness and complexity of the terms employed (such as *fairness, diversity,* and *inclusion*). As their operationalization will always lead to a reduction of their extent, we risk losing or cutting short important societal debates by employing models which appear manageable but have drawbacks, as will be shown in section 3. We deem this human factor to technology an important feature which—although it may soften the mathematical robustness—renders this technology acceptable to society.

2 Quantifying fairness, diversity, and inclusion

The concepts of fairness, diversity and inclusion have many facets/dimensions. Each of the numerous definitions concentrates on itself a specific context, leading each to different meanings and nuances, in turn depending in complex ways on the situation considered. This may lead to conflicting definitions,

*Corresponding Author.
[1]See [7], p. 15.
[2]See e.g. https://datascience.columbia.edu/news/2020/trustworthy-ai/ and https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai.
[3]See [8].

Proceedings of the 5th Northern Lights Deep Learning Conference (NLDL), PMLR 233, 2024.

despite each describes an equally reasonable situation or position, a typical problem for moral and ethical topics [10], which are characteristically complex and multi-dimensional.

In this section, we discuss two well-formulated approaches to fairness, one coined as an algorithm with fairness constraints [11], the other one [12] defines a score (in pseudocode) for the notion of *diversity*, arguably the most discussed positive term for a fair algorithm to assist human decision making. Both examples focus on a seemingly simple application, namely the output of search engines.

Dimensions of diversity. An important point to clarify is the distinction of *heterogeneity* of a data set and the *diversity* it displays, through cues in its items. Checking for diversity is restricting our focus on a subset of cues (often equated with (annotated) attributes in an image or text) which transport meaning related to the dimensions of diversity important to the society in one (geographic or cultural) region at a certain point in time.

There is no universal definition or definitive list of dimensions of diversity accepted worldwide. Besides age, gender/gender identity, and race, other dimensions considered include color, education, ethnicity & national origin, immigration status, income & socioeconomic status, marital status, occupation, parental status, political beliefs, physical & mental ability, religious beliefs, sexual orientation, and veteran status[4]. There may be alterations due to specific focus situations in one country. For example, the Canadian government uses the acronym 2SLGBTQIA+[5] instead of LGBTQIA+ to outline the diversity of sexual orientation and gender identities, adding a facet to value the traditions of indigenous people.

The underlying principle of the lists of diversity dimensions and acronyms may be formulated as to isolate attributes of an individual, which are related to a inequality caused by social structures of power and influence [12]. As such, the notion of diversity varies across geographical locations, and it is dependent on time, the values of the particular society and, pragmatically, on the attention particular cases receive.

In the subsequent discussion, we limit ourselves to a few examples, designed specifically for certain ML use cases.

2.1 Enforcing diversity through fairness constraints

Ranking algorithms are used in search engines, news feeds and recommendation systems. Our example paper [11] gives a solution to the problem that the Hungarian algorithm, which solves the ranking optimization problem as a complete bipartite graph problem, may result in one attribute being overrepresented at the expense of another, by adding constraints to enforce diversity in the ranking.

Definition of fairness. *Fairness* is interpreted in this context as a better representation, e.g the number of items with a certain sensitive attribute l is not allowed to exceed a certain upper bound U_{kl} within the top k positions of the ranking. Solving this constrained ranking problem means that the value w.r.t the original rank quality metric is maximized, while the constraints are respected.

The authors of [11] make the case that enforcing diversity of the output by fairness constraints is more effective than incorporating diversity in the objective function, in particular, as [13] showed that no single diversification function can satisfy a set of natural axioms that one would want a fair ranking to have.

The price to pay is a computational one: the constrained ranking maximization problem is NP-hard, and so is checking if a complete feasible ranking exists. Luckily, in the situation of [11], the precise properties of the classical objective functions in use for the problem allow to construct an algorithm with linear run-time. The algorithm *approximates the constraints*, i.e., violations do occur sometimes.

2.2 Diversity metrics in ML tasks

The subset selection problem is examined in [12]. The prime example of the authors' concepts is a recommendation algorithm for images or movies, i.e. a subset of images or movies selected for a person p performing a query q.

The authors define the *diversity* of an instance as the aggregate statistics of the relevant attributes[6] in the instance. This is quantified via a function of the *presence score* of an attribute a.

Tracking back the author's approach, we may formalize: the user is characterized by a string of relevant attributes such as *Gender:Woman, Skin:FitzpatrickType 6, Age:70*. The binary attribute function a takes as inputs a user (person/individual) p, the subject of the query, and returns 1 if individual p has attribute a, and zero otherwise. In analogy to that, we formalize the following sentence of the authors "define $a(Z_q)$ as a function of a within Z_q, such as the proportion of instances $x_q \in Z_q$ that contain a": Setting N the number of instances in Z_q, and assuming each instance contains at most one relevant cue $x_{q,i}$ related to the attribute $a \in A$,

[4]These 16 dimensions of diversity are listed on https://www.aauw.org/resources/member/governance-tools/dei-toolkit/dimensions-of-diversity/.

[5]See https://women-gender-equality.canada.ca/en/free-to-be-me/2slgbtqi-plus-glossary.html.

[6]With "relevant attributes" we abbreviate the definition given on [12], p.118, which reads: "A [set of attributes] is defined in light of human attributes involved in social power differentials, such as gender, race, color, or creed".

we have the binary variable "presence", also denoted by a, and defined

$$a : Z_q \to [0,1] \quad (1)$$
$$x_{q,i} \mapsto a(x_{q,i}) \quad (2)$$

and the accumulated presence of attribute a in Z_q reads $a(Z_q) = \frac{1}{N}\sum_{i=1}^{N} a(x_{q,i})$.

To check for diversity, set a target lower bound on the presence of an attribute through cues in the subset with the value $l_a \in (0,1)$ and an upper bound with value $u_a \in (0,1)$, and $l_a < u_a$. The diversity score should return zero if any attribute is under- or over-represented in a subset X_q, which is why the authors define the presence score for a fixed attribute attribute $a \in A$ as

$$Presence_a(X_q) = f(a(X_q), l_a, u_a) \quad (3)$$

with f a function being zero outside the admissible interval $[l_a, u_a]$ and monotone increasing inside $[l_a, u_a]$. The diversity score is the the composition with an aggregation function g over all attributes $a \in A$, i.e. $Diversity_A(X_q) = g\Big(Presence_a(x_q)\Big)$, where g can be set by the designer of the scoring system as the minimum, maximum, or average presence score.

Discussion. While the concept of the diversity score certainly has its merits, its operationalization is not completely laid out by the authors of [12], and several issues arise. Firstly, a programmer wishing to implement the diversity score will be frustrated by the sloppiness of notation and the usage of synonyms, the constant switch between instances as being defined as one image or a set of recommended movies with a multitude of relevant items in each of them, or the inconsistencies of the presence function of an attribute being applied to an item in an instance, an instance consisting of one image or an instance consisting of a set of recommended images, which raises the question on possible adaptations of the function as taking values in the integers instead of being binary. But also forgetting such details, the ideas are not fully developed: first, while being well-meant, cues that act as proxies to an attribute may easily be stereotyping, i.e. labeling cues such as 'pink lipstick' or 'high heels' to act as proxies for the *Gender:Woman* attribute.

Second, the implementation of the diversity score requires the creation of an exhaustive, scaled (indexed) list of attributes, for which the *presence score* needs to be calculated. Here, the programmers again need to make a choice themselves, which can strongly influence the score, e.g. on whether to use two, three or more attributes for gender identities.[7]

For some diversity dimensions, a list of unanimous attributes may not be available, e.g. for race or color. Indeed, the commonly used Fitzpatrick skin type (FST) has become a proxy for race, though it was neither designed for this usecase, nor is it adapted to it: The FST was originally developed for white people, on the basis of how they react to sun exposure, and its correlation to constitutive skin color is quite poor [15]. Also, self-identification can differ dramatically from the assigned values, e.g. Japanese women often self-identify as FST type II, though there is only type V reserved for Asian skin. A dermatologist even speculated "the true number of skin colors is unknown but likely is infinite"[8]. Despite the technical restriction that, for a score along a list of to be feasible, the list of attributes needs to be finite, there is also a lower bound on the fine-grainedness of a meaningful scale of skin colors, which may appear lighter or darker, depending on illumination, makeup, and many more factors[9].

2.3 The inclusion score

Getting the diversity score right may not lead necessarily to a balanced data set in all respects. E.g. despite a 50:50 share of pictures with a cue for the woman attribute and the man attribute, respectively, the data set could still consist of pictures displaying each gender in a stereotypical situation. The interesting take on this problem, as suggested by [12], is to define not only a diversity score, but also an inclusion score, which captures if an individual is well-represented by the returned subset selected.

By definition, 'inclusion' has a reference individual or characteristic group, e.g. an action or a subset of items is inclusive *with respect to* a particular reference group. The inclusion score of an item $x_{q,i}$ along one attribute is defined in [12] as a function of the representativeness value $rep_a(i,p,q)$ and the relevance (denoted by $rel(x_{q,i}) = rel(q,i) \in [0,1]$), where the latter is simply reporting how well the output of the (recommendation) algorithm answered the query of the user.

$$Inc_a(x_q, p, q) = f(x_{q,i}, rel(q,i), rep_a(i,p,q)) \quad (4)$$

"Relevance" per se is not a measure of the (non-) discriminatory functioning of the algorithm, and was set to 1 by the authors in all examples[10], therefore we omit it in the subsequent discussion.

As no concrete example or suggestion for the function f in the inclusion score was given, we report the

[7] Currently, approximately 30 different gender identities are described in [14], many of which will be extremely underrepresented in the available image data sets, leading to a failure of classification algorithms to correctly identify them.

[8] See [15], page 78.

[9] The reliable detection of an attribute in a set of images with heterogeneous quality is a different, technical question, which, we don't discuss here.

[10] Indeed, in [12], no further information was given on when the relevance score should be a real-valued number smaller than 1.

conditions mentioned by the authors within their text:

- $Inc_a(x_q, p, q) \in [-1, 1]$, so $Inc_a < 0$ means representation of the opposite and $Inc \approx -1$, means the instance is *stereotypical*.
- If $Inc \approx 1$, means p's attributes are well aligned in the instance.
- $Inc \approx 0$ means that x_q contains few or no items with attributes aligning with p.
- If each instance contains only one relevant item, then $Inc_a = rep_a$ the representativeness of the item.
- If there are many relevant items in the instance, f might measure the median representativeness of some items in the instance.

Discussion. A conceptual question mark in the above definition of inclusiveness is whether such a 'score' is defining a metric (a universal mathematical object): the representativeness score is dependent on the subject attributes to which the comparison is made. In the running example of image or movie recommendations, known user attributes means that all attributes of the individual conducting the query have to be revealed. Inclusiveness in this restricted sense then conflicts with privacy.

If the subject is not known to the algorithm a priori, or should not be known for privacy reasons, the score has to be computed for all possible combinations of attributes[11], which has a terrible complexity. Moreover, this is rather inflexible, as the calculation has to be re-run if the list of attributes is changed. But even if computational problems were to be solved, two issues remain:

First, whether a cue-based score is necessarily referring to stereotypes, and therefore non-inclusive towards individual preferences outside of the majority? Recalling the high heels example, women wearing high heels may feel included by this cue, but those who don't like high heels may not feel included.

The second issue is the problem of representation of very small minorities, of which few items with relevant cues exist in the data set. If those are not relevant to the query, the representativeness cannot be high enough to reach a good inclusion score.

3 The gap between model and reality

The technical approach to concepts like diversity and inclusion demonstrates that sophisticated models - even with best intentions - fail to rebuild these blurry concepts for a model world. This translation process, however, is key to an AI system that meets human-like requirements for these ethical demands. Criticized as 'solutionism' [16], the outcomes of such translations of qualitative aspects (or demands) of the 'real' world into computeable models will suffer from what is normally the greatest asset of a model: a complex issue is narrowed down to its relevant parts to make it processable in specific contexts.

With the concepts of 'diversity' and 'inclusion'—and we can add the more prominent example of 'fairness' as well—aspects of the 'real' world are to be modelled without having a positive, universal example of them [17]. These aspects usually enter the stage of (public) awareness when they are found to be missing. Thus, rather than being concepts with strong definitions, they represent ideas of how the world should be [18].

While this might already be a problem in human-to-human interaction (two people employ the same term but do not agree on its meaning), translating these abstract concepts into scores to produce measureable and computable models adds a new dimension to the problems caused by these terms' blurriness: The designers of such systems may (honestly) say that these concepts have been addressed, yet in the AI system their interpretation is 'one-dimensional', i.e. only one out of many possible meanings of these concepts are taken into account. This might unintentionally narrow down the broad range of meaning of these terms, which can lead to less intense debates on these problems and eventually to a data-driven strong definition of these concepts. The price of which, however, would be high: the multi-dimensionality and therefore the fuel for rich debates on important problems would be lost[17].

3.1 Examples of ethical solutionism in AI

Now, one could argue that with many different systems from different designers this problem become obsolete, as there will be a selection of different approaches implemented. Yet, the history of ethics shows that once a concept has been proven useful, it can no longer be stopped: in biomedical ethics, Beauchamp's and Childress's *Principlism* [19] has become quasi-standard and in the debate on self-driving cars [20], utilitarianism dominates. Each of these approaches to ethics is neither wrong nor misleading, on the contrary, it is their fitness for their purposes which has become the foundation of each's success. This success, nevertheless, is paid for by a reduction of opportunities to think outside the box of standards. Utilitarianism, for example, has its merits in technological contexts as it is an approach which involves calculations [21]. Yet, other aspects of ethics (e.g. duties against oneself or others) are neglected. This may repeat in ML-based systems,

[11] The output may be given either as a vector, listing all scores an a predefined order, or as an average score of some kind.

where this one-dimensionalization will additionally be invisible, as the way issues such as fairness, diversity, or inclusion are dealt with are not explicitly spelled out, but remain hidden within the system.

3.2 Technical approximations of ethical concepts

Another problem connected to a hidden algorithm dealing with unsolved real-world issues is that the technical 'solution' can easily be mistaken as a real one. This might result in less attention to these issues in the real-world, caused by the model which was meant to 'serve' the real world. It is this (until today) unbridgeable gap between model and reality which forces us to reconsider technical solutions that seem advantageous at first glance.

In other words, the algorithmization of ethical concepts, exemplified above, hides unsolved problems of humanity in a system that could be mistaken as the 'solution' to these problems. As a more well-known analogy, jurisprudence and law practice in courts may aim to establish justice, but we cannot equate justice with law. However, we may consider the law as an "approximation" of the idea of justice.

The approximate solution may suffice in many cases, but may fail to provide an acceptable solution in some cases. Just as a court's decision may fail to establish justice in some cases, an algorithmic solution to diversity may fail in a particular test case. In the case of jurisprudence, such instances are mitigated by an appeal to the next higher court instance, and in the case of algorithmic solutions to ethical concepts, human oversight needs to involved to allow for an adequate followup.

4 Conclusion

Technical feasibility does not guarantee an outcome that can be wanted by all. Unsolved issues in interaction between humans will not suddenly disappear when this problem is handed over to machines. However, this does not mean that such endeavors are futile.

Every model-based approach to reality—especially to fix reality—will fail to that extent to which we have no positive idea of how the world we live in should ideally be like. We know what we *do not* want but we cannot exhaustively, let alone consistently express what we *do* want. Thus, caution is needed when systems declare to do what humanity hasn't achieved yet. Such approaches can be helpful reminders of our imperfectness. But they should not be used to hide our failures, or as easy ways out of the complexity of the world by reducing it to a model we think we can manage.

5 Acknowledgments

The authors thank the anonymous reviewers for their valuable feedback, which helped improve this paper. Both authors would like to thank the Federal Ministry of Education and Research of Germany for funding the research centers MCML in Munich and ScaDS.AI in Dresden and Leipzig. Diebel-Fischer also thanks the Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus for the financial support for ScaDS.AI.

References

[1] A. Jobin, M. Ienca, and E. Vayena. "The global landscape of AI ethics guidelines". In: *Nature machine intelligence* 1.9 (2019), pp. 389–399.

[2] T. Hagendorff. "The ethics of AI ethics: An evaluation of guidelines". In: *Minds and machines* 30.1 (2020), pp. 99–120.

[3] D. Lauer. "You cannot have AI ethics without ethics". In: *AI and Ethics* 1.1 (2021), pp. 21–25.

[4] L. Munn. "The uselessness of AI ethics". In: *AI and Ethics* 3.3 (2023), pp. 869–877.

[5] J. Morley, A. Elhalal, F. Garcia, L. Kinsey, J. Mökander, and L. Floridi. "Ethics as a service: a pragmatic operationalisation of AI ethics". In: *Minds and Machines* 31.2 (2021), pp. 239–256.

[6] T. A. Griffin, B. P. Green, and J. V. Welie. "The ethical agency of AI developers". In: *AI and Ethics* (2023), pp. 1–10.

[7] A. Schmitz, M. Akila, D. Hecker, M. Poretschkin, and S. Wrobel. "The why and how of trustworthy AI: An approach for systematic quality assurance when working with ML components". In: *at-Automatisierungstechnik* 70.9 (2022), pp. 793–804.

[8] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton. "Saving face: Investigating the ethical concerns of facial recognition auditing". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 145–151.

[9] B. Ammanath. *Trustworthy AI: a business guide for navigating trust and ethics in AI*. John Wiley & Sons, 2022.

[10] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini. "A clarification of the nuances in the fairness metrics landscape". In: *Scientific Reports* 12.1 (2022), p. 4209.

[11] L. E. Celis, D. Straszak, and N. K. Vishnoi. "Ranking with fairness constraints". In: *arXiv preprint arXiv:1704.06840* (2017).

[12] M. Mitchell, D. Baker, N. Moorosi, E. Denton, B. Hutchinson, A. Hanna, T. Gebru, and J. Morgenstern. "Diversity and inclusion metrics in subset selection". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 117–123.

[13] S. Gollapudi and A. Sharma. "An axiomatic approach for result diversification". In: *Proceedings of the 18th international conference on World wide web*. 2009, pp. 381–390.

[14] Wikipedia contributors. "List of gender identities". In: *Wikipedia* (Aug. 2023). URL: https://en.wikipedia.org/wiki/List_of_gender_identities.

[15] O. R. Ware, J. E. Dawson, M. M. Shinohara, and S. C. Taylor. "Racial limitations of Fitzpatrick skin type". In: *Cutis* 105.2 (2020), pp. 77–80.

[16] E. Morozov. *To save everything, click here: The folly of technological solutionism*. PublicAffairs, 2013.

[17] H. Diebel-Fischer. "Brave New Model World? The Problem of a Data-Driven Construction of a Better (Model) World. Budapest Workshop on Philosophy of Technology 2023". [forthcoming].

[18] H. Diebel-Fischer. "Technisch realisierte Ethik? Anthropologische Perspektiven auf das Verhältnis von Technik und Ethik". In: *Mensch und Maschine im Zeitalter 'Künstlicher Intelligenz'. Theologisch-ethische Herausforderungen*. LIT, 2023, pp. 49–61.

[19] T. L. Beauchamp and J. F. Childress. *Principles of biomedical ethics, 8th ed.* Oxford University Press, USA, 2019.

[20] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan. "The moral machine experiment". In: *Nature* 563.7729 (2018), pp. 59–64.

[21] J. H. Bentham Jeremy Burns (ed.) *An introduction to the principles of morals and legislation*. Oxford: Clarendon Press, 1996.