Can Justice be a measurable value for AI? Proposed evaluation of the relationship between NLP models and principles of Justice

Lidia Marassi[1], Narendra Patwardhan[1] and Francesco Gargiulo[2]

[1] *University of Naples Federico II, Department of Electrical Engineering and Information Technology, Naples, Italy*
[2] *Institute for High Performance Computing and Networking of National Research Council, ICAR-CNR, Naples, Italy*

Abstract

NLP models, such as chat-based generative ones, have gained widespread use and influence in various fields. Considering AI as a valuable resource, it seems critical to determine the ethical principles to which these models should adhere to be considered usable. Applying the philosophical concept of justice to the evaluation of NLP models can help improve their functioning, performance, and overall ethical implications. Although strict adherence to traditional codes of justice may not be sufficient, it is proposed that the concept be adapted to evaluate the performance of NLP systems. Creating a rating scale based on these values may help estimate the "amount of Justice" or fairness demonstrated by different NLP systems. By emphasizing the importance of certain values central to assessing the fairness of NLP models, it becomes possible to develop parameters for evaluation. As these models become increasingly pervasive, this research can help improve this models' performance and their socio-cultural significance. The awareness of building more equitable technologies raises consciousness about the potential consequences of irresponsible use of these models, highlighting the significance of ethical considerations. This approach can lead to a more comprehensive understanding of these systems, promoting their improvement and fostering greater knowledge of the ethical implications associated with their functioning.

Keywords

Justice, NLP, trustworthy, ethics, fairness, measure

1. Introduction

"What is the right measure of justice?". If this question is the basis of historical philosophical debate, today the issue of determining Justice must come to terms with technological progress. That is, what does justice mean in the field of AI? By what metrics can one assess whether a system is fair or just? More importantly, how should this fairness be measured?

These questions are important not only from a moral point of view in the classical sense, as we consider that different philosophical theories offer different approaches to measurement, but also ethically from a more practical point of view. Artificial intelligence is now within everyone's reach (or nearly so), and technological solutions are part of our everyday lives. NLP models have made massive progress in recent years, finding application in a variety of different fields. In recent times, chat-based generative models have experienced an incredible surge in interest and use thanks to numerous companies releasing their language models for free or open-source (ChatGPT, You Ask, Microsoft Bing are some of the early examples of this phenomenon).

The growing interest in NLP models and their pervasiveness suggest that it would be prudent to evaluate their performance in terms of fairness. The philosophical concept of Justice, in this case, could be applied precisely to the evaluation of these technological solutions. In this paper we propose a possible theoretical basis for the development of a system for evaluating the degree of justice of NLP models. By asking ourselves to consider the "rightness" of these models, we can

HCAI4U 2023: Workshop on User Perspectives in Human-Centred Artificial Intelligence, September 20, 2023, Turin, Italy.
lidia.marassi@unina.it (L.Marassi); narendraprakash.patwardhan@unina.it (N. Patwardhan); francesco.gargiulo@icar.cnr.i (F. Gargiulo)

help not only improve its operation and performance, but also to make society, and AI, more ethical.

2. Justice theories and AI

The topic of Justice is not new to the field of AI. In search of inspiration and practicality, several researchers have turned to theories of distributive justice, and in particular, the theories of John Rawls [1], recently dubbed as "the favorite philosopher of artificial intelligence" [2]. Drawing from the philosophical realm, researchers aim to mitigate social injustices by quantifying, measuring, and redistributing resources among individuals in society [3]. Theories of distributive justice propose that justice can be assessed, at least in part, by examining how benefits and burdens are equitably distributed among individuals in society [4]. These theories are valuable for community members, as researchers and policy makers, as they provide proper guidance for understanding the concept of Justice and determining the relative fairness of different societies or systems. In the field of machine learning, an area called fair machine learning (fair ML) has been emerging with the goal of addressing "algorithmic injustice" or bias, as it seeks to apply theories of distributive justice to the design of machine learning systems [5].

Researchers have also attempted to formalize these theories from a quantitative perspective, focusing on concepts such as equality of opportunity [6] and elements of John Rawls' influential distributive justice theory [7].

Despite the great influence that theories of distributive justice seem to have in the field of AI, here we instead propose an approach closer to the so-called Capability Approach. As argued by Amartya Sen [8], Rawls' measure of justice would indeed seem to prove insensitive to the heterogeneities of people and social context.

Especially when applied to the field of AI, distributive justice approach may not consider the different ways in which some people are (or are not) able to convert resources into well-being. It is considered that what usually represents an opportunity for most people, namely converting a resource into a valuable state of being or acting, may be precluded to others because of their individual differences and social context. For example, for a blind person, the purchase of a computer does not result in a means to enjoy the benefits of Internet access if the computer is not compatible with screen reader technology. In this situation, the computer becomes a false opportunity to achieve the corresponding valuable state, which is surfing the Internet.

NLP models, indeed, can be used for text creation (e.g., articles, reports, product descriptions...); however, if the model does not consider cultural sensitivities or lexical appropriateness, it could generate inappropriate or offensive text for certain user groups. This would make the automatic generation resource a false opportunity to create relevant and respectful content for different communities of readers. Capability theorists argue that the real support for well-being lies not in the resources themselves, but in what people are able to achieve through their use. Thus, Justice does not reside in resources, but in people's abilities to use them.

2.1. Ethical principles

For individuals to be able to use resources to improve their own well-being, it seems necessary that access to them be guaranteed by a set of ethical principles. If we consider AI as a resource (to date, perhaps one of the richest), then the question about the principles it should adhere to be usable is indeed fundamental. While sticking strictly to classical codifications of the concept of Justice can lead to a dead end, in this paper, we suggest instead adapting the concept to performance evaluation of NLP systems. Referring to the capability approach theory, it considers several ethical principles [9], among which we consider:

- Freedom: It is argued that Justice should be evaluated since people's ability to realize their own choices. This implies promoting a wide range of freedoms and opportunities to enable individuals to pursue their aspirations.

- Human Dignity: as the inherent dignity of every human being. Justice should ensure that people have the capabilities and have access to the resources they need to live a life of dignity.
- Equality of Opportunity: this principle emphasizes the importance of providing all individuals with equitable access to the opportunities they need to realize their potential.
- Social inclusion: it calls for the removal of discrimination and inequalities that limit participation and inclusion.
- Sustainability: both environmentally and socially. Indeed, justice requires the responsible management of resources and the preservation of opportunities for future generations.

3. Ethical principles as measurable values for NLP models

In the field of AI, with reference to NLP models, there are several values that seem to be significant with respect to this view of Justice. We believe that it would be possible to assess a model's ability to satisfy these principles of justice if they were adapted to NLP models. As chatbots are artificial intelligence systems designed to interact with people, they can have a significant impact on people's lives, both in terms of well-being and justice.

The identification of values, in terms of variables, can be the basis for creating a rating scale to estimate the "amount of justice" (which we will mean in terms of fairness) of different NLP systems. Based on the attribution of a numerical estimate of the NLP model passed under consideration, it will be possible to quantify its "fairness" in practical terms. This result would not only allow for a more in-depth analysis of these systems to work toward improving their performance but would also have significance in socio-cultural terms.

Indeed, awareness of the need to build more equitable technologies seems to be the basis for increased awareness of the possible consequences of ill-considered use of these models, especially if they took ethical aspects into account. With this approach, an in-depth and quantitative evaluation of each parameter considered in NLP models is proposed, helping to understand how the values identified affect the fairness of the model.

The process proposed in the research involves:

(A) Identification of Reference Values and Contextualization in NLP Systems:

- Identifying the value (e.g., fairness): understanding what this value represents within NLP systems (e.g., fairness as fairness of treatment);
- Relevance of Value to Model Fairness: understand why the identified value is relevant to greater fairness in the model (e.g., a fair model does not discriminate and provides accurate and relevant results for all users);
- Choice of Evaluation Metrics: identify suitable metrics for quantitative measurement of the identified value, also drawing on relevant scientific literature (e.g., to assess fairness, an important metric might be the degree to which the model is able to identify personal information regarding the interlocutor, such as gender, after interaction with the algorithm. A fair model should avoid making assumptions or discriminating based on these personal characteristics. Thus, the more the model can predict who it is talking to, the less fair it would be considered).

(B) Quantitative Assessment of Model Fairness:

- Measurement of Identified Values: measuring the values identified in the performance of the NLP model;
- Assigning Numerical Scores: assigning a numerical score to each evaluated value, representing the amount of fairness present in each aspect considered;
- Creating a Graph: representing the measured values through a graph that will show the amount of equity for the different aspects considered.

In the end, a graph with the measured values will be obtained, assigning a number to each evaluated value. On this basis, it will be possible to have a quantitative measure of the fairness of the model, differentiated for the different aspects considered. Here, we will focus on explaining the importance of certain values, central to the evaluation of the fairness of NLP models, which can later be developed in terms of parameters to be evaluated.

3.1. Transparency – Freedom

Transparency is a fundamental pillar in evaluating the justice of NLP systems. Transparency in AI is often used interchangeably with explainable AI, but it's more focused on ensuring that a model is open and visible [10]. Users must have the ability to understand how these systems make decisions and process their data. In the Capabilities Approach, freedom of choice is considered a fundamental element of human well-being. It is not enough to have different options available, but it is essential that people have a real opportunity to understand and evaluate these options so that they can make informed and well-informed choices. This often implies the need for access to clear, understandable and transparent information about the options available and the consequences of the choices that can be made.

Providing access to model and training data internals enables users to examine the inner workings of NLP systems and ensures accountability. Model cards provide detailed information about the structure, design choices, and components of the system. They offer insights into the underlying mechanisms of the model, such as the types of neural network layers used, attention mechanisms, and pre-processing steps. Understanding the model architecture empowers users to assess the strengths, limitations, and potential biases of the system.

Users should also have the right to remove their personal data from the training corpus, respecting their privacy and building trust. Accommodating data removal requests is essential for ethical practices.

Additionally, detailed algorithm documentation empowers users with a deeper understanding of decision-making, data processing, and output generation. This knowledge helps users assess the system's reliability, consistency, and potential vulnerabilities.

This connection is important because if an NLP model is opaque and does not provide explanations or justifications for its predictions, users' freedom of choice could be compromised because they would not understand how decisions were made and what options were considered. When users understand how a model makes decisions and operates on their data, they can assess whether the model meets their values, expectations, and preferences. This awareness enables users to make informed decisions about whether to trust the model.

Greater transparency in NLP models helps to ensure that users could make informed and well-informed choices by having a better understanding of how the model works and how it reaches its conclusions. This is especially important in critical applications where model decisions can have a significant impact on people's lives, such as in medical, legal or public policy decisions. In addition, greater transparency can help identify and address any bias or prejudice in the model, promoting greater equity and inclusion in NLP applications. Additionally, transparency facilitates the identification and mitigation of biases and discriminatory tendencies, allowing for continuous improvement and the establishment of trust between users and developers.

3.2. Fairness of treatment – Equality of Opportunity

Despite the remarkable accuracy of AI solutions in various applications, learning algorithms can still rely on social biases encoded in training data to make predictions. This can happen even when information about gender and ethnicity is not explicitly provided to the system. Therefore, machine learning algorithms potentially risk encouraging unfair and discriminatory decision making. Fairness of treatment, in the context of a natural language processing system, refers to the system's ability to treat the users involved in the interaction fairly and impartially. Indeed, the system should not discriminate against or unfairly favour any individual because of personal characteristics such as gender, ethnicity, age, or sexual orientation. This value relates directly to the fundamental principle of equality of opportunity and the prohibition of discrimination. An NLP model that promotes fair treatment will have to provide homogeneous and impartial responses, regardless of their individual user characteristics. The system that meets this parameter must avoid stereotyping, bias, or discriminatory treatment in processing users' requests.

Measuring fairness of treatment in an NLP system can be problematic, involving analysing training data, identifying any biases embedded in the model and implementing techniques to mitigate them. Careful consideration must be given to data input, evaluation metrics, and mitigation strategies to ensure that the system maintains fair and non-discriminatory treatment of all users. Assessing bias is critical to better understanding and addressing unfairness in NLP models. This is often done through equity metrics [11] which quantify differences in the behaviour of a model across a range of demographic groups. Including the fair treatment parameter in an NLP system is intended to ensure that all people have equal access, opportunity and fair treatment while using the system, to incentivize an inclusive and unbiased experience for users.

3.3. Accessibility – Human Dignity

The purpose of accessibility is to remove any barriers that might prevent certain people from accessing information and interacting with the system efficiently. In the context of an NLP system, the system should be designed to be usable by people with visual or hearing disabilities. This may require the implementation of text-to-speech capabilities or support for reading text by assistive devices.

Also, the system ought to be designed in a way that allows people with motor or speech disabilities to interact with it effectively (such as using voice commands or alternative interfaces). Human dignity is considered as a fundamental principle that recognizes the inherent worth and dignity of every individual, regardless of their personal characteristics.

An accessible NLP system respects human dignity by ensuring that all people have an equal opportunity to access and interact with the system. If an NLP system is not accessible, it may discriminately treat some people or exclude them from accessing information or services provided by the system. This can be harmful to the dignity of the individuals involved, resulting in a sense of marginalization, discrimination, or disadvantage.

While training multimodal architectures can be costly due to the need of balanced data between modalities, text based pretrained models can be extended for accessibility. Parameter-efficient transfer learning techniques can be employed to adapt pretrained models to specific accessibility requirements. Fine-tuning models on specialized datasets that capture the nuances of accessibility-related interactions can improve the system's performance and responsiveness to diverse user needs. Pretrained models that support different modalities can be chained, such as Whisper [12], a speech-to-text model to a text-to-text model, to extend their capabilities.

Accessibility is also connected to the value of social inclusion as the main aim of social inclusion is to guarantee that all people, regardless of their differences and skills, have the chance to participate fully and meaningfully in society. This encompasses access to resources, services, opportunities and basic rights. Accessibility is a crucial aspect in the context of NLP systems; proper attention to accessibility can ensure that the system can be used by a wide range of users and that no one is marginalized or disadvantaged because of limitations or disabilities.

3.4. Equity in access – Social Inclusion

Closely related to the value of accessibility, we consider equity in access. By this expression, we refer to the principle that ensures equal opportunity to use NLP models, regardless of the differences, abilities, or socio-cultural backgrounds of users.

To promote equity in access, the system should be designed and implemented in such a way that it is accessible to people with different levels of technological competence or language proficiency. Another important aspect of equity in access is to address inequalities in access caused by socioeconomic or geographic factors, to ensure that people from disadvantaged groups or marginalized communities also could access NLP systems. As social inclusion refers to the process of engaging and participating all people in society, equitable access to NLP systems is a key element in fostering social inclusion. When NLP systems are designed with equity of access in mind, barriers that may marginalize certain individuals or groups are reduced, thus promoting a more inclusive social environment and technology space.

3.5. Long-term utility – Sustainability

In this context, the principle of sustainability ties in with justice insofar as it aims to ensure equitable access and benefits for all stakeholders, both in the present and in future generations.NLP models can indeed be effective and relevant over time without becoming obsolete or less useful. To be sustainable over the long term, they must be able to maintain their accuracy, understanding, and language-generating capacity even in the face of new scenarios and contexts (such as evolving languages, changes in user needs).

NLP models, to meet this value, should be designed with their possible flexibility and adaptability to context in mind, thus avoiding model obsolescence. For example, considering adaptation to change in long-term utility for NLP systems requires a strategic and flexible approach throughout the life cycle of the system [13]. It is essential to continuously monitor the operation of NLP systems and evaluate performance against relevant metrics. This enables early identification of problems or changes in the context and user needs.

To ensure equity over the long term, it is important to proactively explore the system and continuously assess its relevance and impact on different populations. This could include close monitoring of the model's performance on different demographic groups and early identification of any disparities or biases in the model's operation. If the model shows unequal or discriminatory performance on certain groups, corrective measures should be taken to improve its fairness. Thus, long-term usefulness is an important aspect of sustainability in NLP model, as ensuring the long-term usefulness of NLP models means developing models that can both remain relevant and helpful over time, thus allowing equitable access to advanced language technologies to be preserved for future generations.

4. Conclusions

The aim of this position paper was to propose a possible theoretical basis for the development of a system for evaluating the degree of justice of NLP models. Here, we have brought attention to the need to assess the fairness of these algorithms more reliably by analyzing what could be the core values in evaluating their adherence to the philosophical principles of Justice. This emphasizes the importance of fuzzy evaluation based on ethical and moral principles in AI development.

This work is part of the Human Centered Artificial Intelligence Master (HCAIM) program, sponsored by the University of Naples Federico II and in collaboration with the National Research Council (CNR) of Naples. The HCAIM program involves four prominent European universities and specialists from various fields to provide students with a comprehensive and interdisciplinary education on human-centered AI. The research aims to create an evaluation system that can classify AI solutions' adherence to philosophical principles of Justice. It seeks to develop socially responsible, technically sound, and ethically acceptable NLP models by considering values like freedom, human dignity, equal opportunity, social inclusion, and sustainability. The proposed evaluation framework provides a basis for assessing the Justice of these models and guiding their improvement.

References

[1] Rawls, John. A Theory of Justice. Belknap Press of Harvard University Press, Cambridge, Mass., 1971. ISBN 978-0-674-04260-5. http://site.ebrary.com/id/10318418.

[2] Procaccia, Ariel. AI Researchers Are Pushing Bias Out of Algorithms. Bloomberg Opinion, March 2019. https://www.bloomberg.com/opinion/articles/2019-03-07/ ai-researchers-are-pushing-bias-out-of-algorithms.

[3] Hoffmann, Anna Lauren. Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse. Information, Communication & Society, 22(7):900–915, 2019. https://doi.org/10.1080/ 1369118X.2019.1573912.

[4] Lundgard, Alan. Measuring justice in machine learning. 2020. DOI.org (Datacite), https://doi.org/10.48550/ARXIV.2009.10050.

[5] Binns, Reuben. Fairness in Machine Learning: Lessons from Political Philosophy. page 11, 2018. http://proceedings.mlr.press/v81/binns18a.html.

[6] Hardt, Moritz, Eric Price, ecprice, and Nati Srebro. Equality of Opportunity in Supervised Learning. Advances in Neural Information Processing Systems 29, pages 3315–3323, 2016. http://papers.nips. cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf

[7] I Hashimoto, Tatsunori, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness Without De mographics in Repeated Loss Minimization. In International Conference on Machine Learning, pages 1929–1938, July 2018. http://proceedings.mlr.press/v80/hashimoto18a.html.

[8] Sen, Amartya. Commodities and Capabilities. North-Holland, Amsterdam, 1985. https://scholar. harvard.edu/sen/publications/commodities-and-capabilities.

[9] Alkire, S. and Deneulin, S., 2009, “The Human Development and Capability Approach”, in Shahani and Deneulin (eds.), An Introduction to the Human Development and Capability Approach: Freedom and Agency, London: Earthscan.

[10] Santosh, K. C., e Casey Wall. AI, Ethical Issues and Explainability -- Applied Biometrics. Springer, 2022.

[11] Paula Czarnowska, Yogarshi Vyas, Kashif Shah; Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. Transactions of the Association for Computational Linguistics 2021; 9 1249–1267. doi: https://doi.org/10.1162/tacl_a_00425.

[12] Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." arXiv preprint arXiv:2212.04356 (2022).

[13] Liu, Huiting, et al. Model Stability with Continuous Data Updates. 2022. DOI.org (Datacite), https://doi.org/10.48550/ARXIV.2201.05692.