AI Orthopraxy: Towards a Framework for That Promotes Fairness

Pablo Rivas, *Senior IEEE*
Department of Computer Science
School of Engineering and Computer Science
Baylor University
Waco, Texas, USA
Pablo_Rivas@Baylor.edu

***Abstract*—This paper introduces the term AI Orthopraxy as the correct practice of AI and a framework that aims to unify some aspects associated with AI ethics. These include standards, legal, and measures of fairness. We draw from existing tools that have been peer-reviewed by academics and discussed in recent literature to provide a mechanism for assessing the level by which a model or AI technology follows the correct practices of ethical AI. This paper describes a preliminary, ongoing, study and shows the early stages of a prototype framework, including a visual representation of the level of AI Orthopraxy of a model using hive plots. This work can potentially help create fair and trustworthy AI built upon the core tenets of accountability, transparency, and fairness. One of the current limitations is that it requires validation of peers that are willing, able, and trained to evaluate an AI model or technology using standards and other novel frameworks.**

Keywords—AI orthopraxy, AI ethics, fairness, transparency, accountability, ethical frameworks, standards, P7000, assessment tools

I. Introduction

Fairness has been analyzed in different contexts relevant to society, such as law [1], justice [2], policy-making [3], moral philosophy [4], computing technology [5], and others. The latter has received significant interest given the rise, resilience, and ubiquitousness of computing technology [6-9]. Data-fueled artificial intelligence (AI) systems have recently gained much attention regarding ethical issues that have come to light [10]. Researches in the field have addressed several ethical issues associated with AI in the form of ethical principles application [11], analysis of classic ethical dilemmas [12], and the proposition of guidelines and standards [13]. However, as AI continues to grow, becoming more accessible to people, and as it expands its applicability, applying a generalized notion of fairness in all contexts has become more difficult. Different criteria for detecting and addressing bias have been proposed to address this problem [14], successfully preventing forms of discrimination [15], leading to fairness [16]. Nonetheless, more forms of bias continue to surface and leading to discrimination and unfairness either caused directly by unforeseen effects in AI models, careless data treatment, or malicious human involvement.

This short paper discloses our current research efforts in pursuit of a standardized AI orthopraxy model that promotes and measures fairness in research and development. This research will study the different parts of existing methodologies that address and promote fairness in AI and closely related fields. These include a) gathering resources from a legal standpoint to determine what is necessary according to the US and international law; b) gathering the existing applicable standards by institutions such as ISO, IEEE, or ACM to determine what *shall and must* be done to comply with ethical AI standards; and c) determine, using the former elements, what are both the minimum requirements and actions (non-ideal) and the best practices (ideal) that need to be followed and carried in order demonstrate a fair AI modeling and mindset.

II. Objectives

The main objectives of the research we are embarking on are:

1) To study legal standard-based frameworks for AI fairness.

2) To determine qualitative and quantitative elements that can objectively measure AI fairness.

3) To propose models for the evaluation of AI algorithms and AI-based technology.

4) To create a self-assessment tool for AI researchers and training modules for AI educators.

The work we are beginning aims to bring together recent advances in legal practice and ethical frameworks or standards informing scientists and technologists about the critical aspects necessary for AI fairness. This work will take advantage of novel taxonomies of AI fairness and create a framework for AI orthopraxy that can potentially be self-sustainable. The potential for self-sustainability can be achieved via ongoing self-assessment and peer-review of AI research and technology under the framework that will be proposed. The success in adopting this framework will be determined by an experiment with AI researchers publishing at major AI conferences, which will make a self-assessment and peer-review colleagues voluntarily.

The dissemination of the framework and the report of the results will advance the field of AI fairness by revealing the interest of scientists to meet the minimum standards or best practices. Furthermore, AI educators and other public and private institutions will have access to the suggested curriculum for training existing and forming AI researchers.

III. Other Similar Efforts

A recent study has made a great effort to categorize past and current efforts in AI ethics [16]. The study highlights the need for research like ours and showcases the idea of an ethic of AI ethics. We propose something similar that has an actionable aspect to the correct (*ortho*) practice (*praxis*) of AI, which we hope people call *AI Orthopraxy*. The term orthopraxis has been typically used in religious literature; however, the term means "correct practice," which is precisely what we want to convey in our research.

We will now examine the current efforts across the different dimensions we want to tie together into an AI orthopraxy framework.

A. Legal Frameworks

Legal frameworks can be challenging to establish into international law and typically require significant partnerships among the nations; however, in the case of AI, given that the producers of AI and the consumers of AI might not always be represented appropriately, legal frameworks have been left to the companies as a matter of self-policing [16-17]. However, such efforts often fail to prevent issues with accountability, transparency, or clear and legally binding self-regulation [18]. However, other types of technological advances, such as the robotics industry, for example, the authors in [19] have made significant progress. The past successes suggest that legal frameworks and legally binding agreements are often born from guidelines and standards, which is what we discuss next.

B. Standards

The IEEE and its members have organized and promoted efforts to guide researchers, practitioners, and developers to design ethical AI [20]. Most recently, the P7000 series of standards is leading the charge to develop standards that have the potential to be at best adopted into legal frameworks, or at worst, be recommended by governments internationally [13]. This is one of the most comprehensive groups of AI ethics standards, which include:

- P7000: Model Process for Addressing Ethical Concerns During System Design.
- P7001: Transparency of Autonomous Systems.
- P7002: Data Privacy Process.
- P7003: Algorithmic Bias Considerations.
- P7004: Standard on Child and Student Data Governance.
- P7005: Standard on Employer Data Governance.
- P7006: Standard on Personal Data AI Agent Working Group.
- P7007: Ontological Standard for Ethically driven Robotics and Automation Systems.
- P7008: Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems.
- P7009: Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems.
- P7010: Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being.
- P7011: Standard for the Process of Identifying & Rating the Trust-worthiness of News Sources.
- P7012: Standard for Machine Readable Personal Privacy Terms.
- P7014: Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems.

While the standards community, comprised of international participants representing many different communities, is the strongest player in AI Orthopraxy, there are other guidelines and general frameworks that have been recently released.

C. General Guidelines or Frameworks

The general frameworks or guidelines that have been found in the literature include a guide for linking basic ethical principles to AI [21] and another that uses a similar approach but is more comprehensive to include a rights-based approach [22]. Another recent attempt to articulate the perspectives of AI ethics has gained attention [23]. The authors certainly provide a big-picture approach to AI ethics; however, it has also received criticism for claiming to be comprehensive while leaving out critical legal arguments [16].

There are two other exciting approaches that have the same practical mindset of what we aim for. The first one is a methodology for documenting datasets rooted in the tenets of accountability and transparency [15]. The authors propose that anyone in the AI community document our datasets, including motivation, composition, collection process, pre-processing, uses, distribution, and maintenance. The premise is that a dataset reported in this fashion will provide the end-user with transparency and also will transfer the responsibility (for accountability purposes) to those who will use the dataset.

The second approach is aimed at AI models [24]. The authors propose using a card where the designers of AI models disclose the following information: model details, intended use, factors, metrics, evaluation data, training data, quantitative analysis, ethical considerations, caveats and recommendations. This also follows a similar philosophy for accountability and transparency.

Our proposed approach follows a similar flavor but adds a particular measure of fairness that is necessary and missing in many current practical approaches.

IV. Proposed Measure of Fairness

For the most part, the pursuit of fairness has been left as a self-regulating item for industry, often used as self-publicity. For

example, Google has the "AI Fairness 360" tool kit and the "What-If Tool"; Microsoft has "fairlern.py"; and Facebook has "Fairness Flow" [14].

Measuring fairness can be too complex. Most academic arguments against a universal measure of fairness can be summarized into the general idea that fairness cannot be applied in general without considering the appropriate context. That is, what might be fair for person P in context X might not be fair to P in context Y. This mindset has hindered progress in this area. But, to this we say 'so what?' The fact that there are no universal approaches to fairness analysis does not mean that we cannot try. We must begin. Thus, we will attempt the following two primary ways to measure fairness: context-dependent and context-free ideas.

A. *Context-free*

One of the most common and recent approaches to measure fairness is through counterfactual theory [25]. For example, counterfactual fairness asks the question: how would a prediction change if a sensitive attribute were different?

This way of measuring changes in the performance of models by altering sensitive features while the general context remains the same is helpful in addressing not only fairness but also bias in general.

B. *Context-dependent*

When the context is necessary for measuring fairness, we can recur to making the optimal decision based on the circumstances for which models were designed to operate. Recent approaches have proposed to solve this problem as an optimization problem with specific constraints that seek fairness in particular contexts [26].

These approaches have an academic flavor that does not focus on particular industries and are simple. While we acknowledge that this list is not comprehensive, we continue to study compatible approaches that are simple, academic, and can be applied by practitioners, designers, academics, and managers. In the next section, we put the pieces together, offering a self-assessment tool for AI orthopraxy.

V. Proposed Assessment Tool

Our preliminary research considers analyzing the correct practice of AI using four major components that have to be verified and documented by three different AI academics. These three areas are posed as questions that are answered using a Likert scale, as shown in Table I. The Likert scale used for likelihood is: (5) Definitely, (4) Probably, (3) Possibly, (2) Probably Not, and (1) Definitely Not. Similarly, the scale for agreement is: (5) Strongly Agree, (4) Agree, (3) Undecided, (2) Disagree, and (1) Strongly Disagree. The table includes an additional pass-fail measure that is compatible with most AI standards, and can be used, for example, by an internal reviewer before submitting any work for publication or deploying AI models. These Likert scales are well-known, easy to follow, standardized forms of assessing certain non-trivial qualities of a subject. In our case, the assessment of correct AI practices based on the likelihood of complying with a standard, or agreement with a representative qualitative statement.

TABLE I. Measuring AI Orthopraxy that Promotes Fairness

Likert Scale	**Measurement Methodology**		
	Question	***Pass***	***Fail***
Likelihood	Will this model comply with AI ethics standards and local legal regulations?	Above *probably*	Below *possibly*
Agreement	Does the properly filled model card reflect a trustworthy model?	Above *Agree*	Below *Undecided*
Agreement	Does the properly filled datasheet reflect trustworthy, bias-free, data?	Above *Agree*	Below *Undecided*
Likelihood	Will this model be fair to all end-users in all cases?	Above *probably*	Below *possibly*

Fig. 1 depicts an example of a hive plot we intend to use as a reporting tool. Every axis in the hive plot corresponds to a reviewer who is vetting the model in question, in this case, three. The scale of every axis is from 4 (lowest) to 20 (maximum). Every color in the plot corresponds to a particular question in Table I, where the first question corresponds to the outermost ring and the last question to the innermost. The score in the center of the figure describes an average score normalized so that the maximum is ten and the minimum is two. Fig. 2, in comparison, shows how a low score would look for a model or AI technology that does not follow the correct practices.

This assessment tool is in its early stages of development; however, the premise is relatively simple, draw from current practical efforts that have been vetted by academics, IEEE standards (P7000 series), model cards [24], datasheets [15], and fairness analysis [25-26]. Further, the proposed model can be scaled to add more reviewers as needed. We are working to determine a recommended number of reviewers as a function of the number of potential users of the AI technology or model, where the absolute minimum is always three.

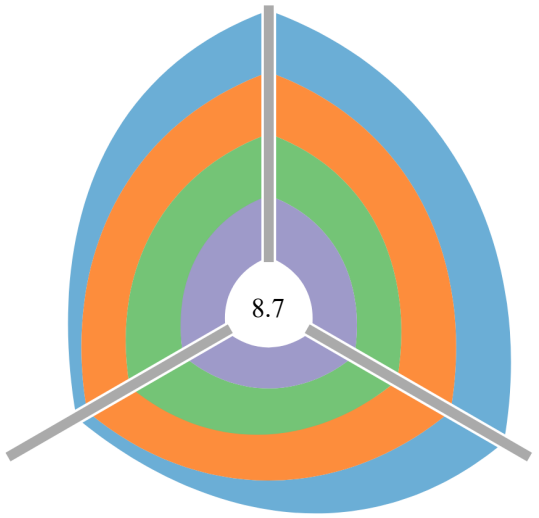

Fig. 1. In this example, the hive plot indicates overall high scores; one of the reviewers reported a low score in the first question.

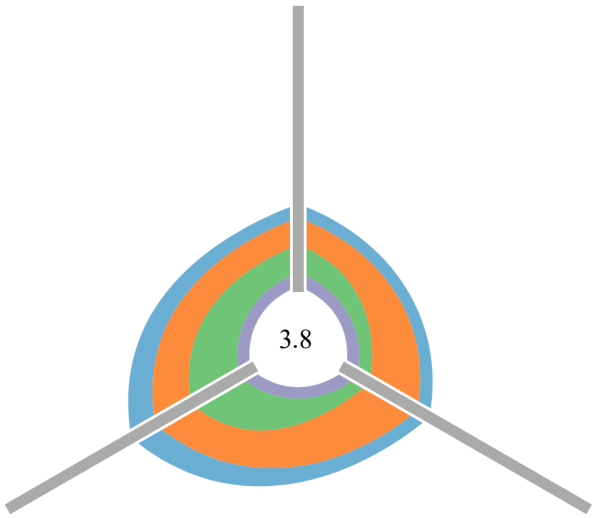

Fig. 2. In this example, the hive plot indicates overall low scores; most of the reviewers reported a low score in the first question and last question.

We are also working on making this tool available as part of our website (https://baylor.ai), aiming to provide a persistent record of the reviews for a particular model in the same way that peer-reviews are done in academic circles. Such records can then be pulled by search engines or indexing services and be displayed alongside searches. However, for now, these are early implementations and studies, and we are looking to drive this further as a short-term goal.

VI. Discussion and Conclusions

The work discussed in this paper aims to bring together recent advances in legal and ethical frameworks or standards informing scientists and technologists about the critical aspects necessary for AI fairness. This work will take advantage of novel taxonomies of AI fairness and create a framework for the correct practice of AI, which we call AI orthopraxy. The work proposed here can potentially be self-sustainable. The potential for self-sustainability can be achieved via ongoing self-assessment and peer-review of AI research and technology under the proposed framework. The success in adopting this framework will be determined by an experiment with AI researchers publishing at major AI conferences, which will make a self-assessment and peer-review colleagues voluntarily. The dissemination of the framework and the report of the results will advance the field of AI fairness by revealing the interest of scientists to meet the minimum standards or best practices. Furthermore, AI educators and other public and private institutions will have access to the suggested curriculum for training existing, and under training, AI researchers which will come at a future date.

The proposed work can empower researchers and technologists to make peer and self-assessments to detect, address, and, therefore, mitigate the risk of unfairness in AI research and development. This work can positively influence the way search is published in the AI field if journal editors and conference program committees promote and recommend authors to voluntarily perform a self-assessment under the proposed framework and, further, if they require this also as part of the peer-review process. Several major conferences are already implementing forms of verification of these aspects, but in the early stages. The voluntary release of these assessments' results can give applied AI technologists confidence that a particular AI algorithm or research was reviewed correctly to mitigate the risk of unfairness if such research is reproduced. For other researchers, this can be an opportunity to highlight research that is safer to the general public than similar algorithms that have the potential for bias that can lead to discrimination if implemented.

Acknowledgment

The author would like to thank the IEEE Standards Association and the P7000, P7003, and P7014 working groups for meaningful discussions and support.

References

[1] L. Kaplow and S. Shavell, "Fairness versus welfare," *Harv. L. Rev.*, vol. 114, p. 961, 2000.

[2] R. Folger and R. Cropanzano, "Fairness theory: Justice as accountability," *Advances in organizational justice*, vol. 1, no. 1-55, p. 12, 2001.

[3] L. Kaplow and S. Shavell, *Fairness versus welfare*. Harvard university press, 2009.

[4] J. Rawls, "Justice as fairness," *The philosophical review*, vol. 67, no. 2, pp. 164–194, 1958.

[5] R. K. Jain, D.-M. W. Chiu, W. R. Haweet al., "A quantitative measure of fairness and discrimination," *Eastern Research Laboratory, Digital Equipment Corporation*, Hudson, MA, 1984.

[6] [dwork2012fairness] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[7] I. Rhee, A. Warrier, M. Aia, J. Min, and M. L. Sichitiu, "Z-mac: a hybrid mac for wirelesssensor networks," *IEEE/ACM Transactions On Networking*, vol. 16, no. 3, pp. 511–524, 2008.

[8] D.-M. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," *Computer Networks and ISDN systems*, vol. 17, no. 1, pp.1–14, 1989.

[9] K. Li and P. Hudak, "Memory coherence in shared virtual memory systems," *ACM Transactionson Computer Systems (TOCS)*, vol. 7, no. 4, pp. 321–359, 1989.

[10] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[11] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang, "Building ethics into ar-tificial intelligence," *in Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 5527–5533.

[12] [awad2018moral] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The moral machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, 2018.

[13] S. Spiekermann, "Ieee p7000—the first global standard process for addressing ethical concerns in system design," *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 1, no. 3, p.159, 2017.

[14] M. Whittaker, K. Crawford, R. Dobbe, G. Fried, E. Kaziunas, V. Mathur, S. M. West, R. Richardson, J. Schultz, and O. Schwartz, *AI now report 2018*. AI Now Institute at New York University New York, 2018.

[15] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daume III, and K. Crawford, "Datasheets for datasets," *arXiv preprint* arXiv:1803.09010, 2018.

[16] T. Hagendorff, "The ethics of ai ethics: An evaluation of guidelines," *Minds and Machines*, pp. 1–22, 2020.

[17] R. Calo, “Artificial intelligence policy: a primer and roadmap,” *UCDL Rev.*, vol. 51, p. 399, 2017.

[18] B. Thuraisingham, “Artificial intelligence and data science governance: Roles and responsibilities at the c-level and the board,” *in 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2020, pp. 314–318.

[19] K. Darling, “Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects,” *in Robot law*. Edward Elgar Publishing, 2016.

[20] K. Shahriari and M. Shahriari, “Ieee standard review—ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems,” *in 2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*. IEEE, 2017, pp.197–201.

[21] Y. Zeng, E. Lu, and C. Huangfu, “Linking artificial intelligence principles,” *arXiv preprint* arXiv:1812.04814, 2018.

[22] J. Fjeld, H. Hilligoss, N. Achten, M. L. Daniel, J. Feldman, and S. Kagay, “Principled artificial intelligence: A map of ethical and rights-based approaches,” 2019.

[23] A. Jobin, M. Ienca, and E. Vayena, “The global landscape of ai ethics guidelines,” *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[24] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, “Model cards for model reporting,” *in Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.

[25] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel, “Counterfactual fairness in text classification through robustness,” *in Proceedings of the 2019 AAAI/ACM Conferenceon AI, Ethics, and Society*, 2019, pp. 219–226.

[26] Y. Zhang, R. Bellamy, and K. Varshney, “Joint optimization of ai fairness and utility: A human-centered approach,” *in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* 2020, pp. 400–406.