Policy Advice and Best Practices on Bias and Fairness in Artificial Intelligence*

An Extended Abstract

Jose M. Alvarez[1,2,*], Alejandra Bringas-Colmenarejo[3], Alaa Elobaid[4,5], Simone Fabbrizzi[4,6,7], Miriam Fahimi[8], Antonio Ferrara[9,10,11], Siamak Ghodsi[5,6], Carlos Mougan[3], Ioanna Papageorgiou[6], Paula Reyero[12], Mayra Russo[6], Kristen M. Scott[13], Laura State[1,2], Xuan Zhao[14] and Salvatore Ruggieri[2,*]

[1]Scuola Normale Superiore, Pisa, Italy
[2]University of Pisa, Pisa, Italy
[3]University of Southampton, Southampton, UK
[4]CERTH, Thessaloniki, Greece
[5]Free University of Berlin, Berlin, Germany
[6]Leibniz University Hannover, Hannover, Germany
[7]Free University of Bozen-Bolzano, Bolzano, Italy
[8]University of Klagenfurt, Klagenfurt, Austria
[9]GESIS - Leibniz Institute, Mannheim, Germany
[10]TU Graz, Graz, Austria
[11]CENTAI, Turin, Italy
[12]The Open University, Milton Keynes, UK
[13]KU Leuven, Leuven, Belgium
[14]SCHUFA Holding AG, Wiesbaden, Germany

Introduction, motivation, and contributions. With the increasing usage of AI models in our daily lives, concerns have been raised on the negative impact of AI models on individuals and society due to their embedded biases [2]. There is a deep academic and social discussion around the alleged neutrality of these algorithmic systems as more examples confirm that such algorithmic systems are "value-laden in that they create moral consequences, reinforce or undercut ethical principles, or enable or diminish stakeholder rights and dignity" [3]. Included in that discussion is the interdisciplinary and growing field of *fair-AI*. In Álvarez et al. [1], we survey the fair-AI state-of-the-art of methods and resources as well as the latest policies on bias in AI, in turn, providing the much needed bird's-eye view for all stakeholders. Further, by

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany
*This is an extended abstract: see Álvarez et al. [1] for the full paper; click here for the open access link.
*Corresponding authors.
jose.alvarez@sns.it (J. M. Alvarez); salvatore.ruggieri@unipi.it (S. Ruggieri)
0000-0001-9412-9013 (J. M. Alvarez); 0000-0002-7968-9853 (A. Bringas-Colmenarejo); 0009-0009-2399-9754 (A. Elobaid); 0000-0003-4374-5806 (S. Fabbrizzi); 0000-0002-0619-3160 (M. Fahimi); 0000-0002-0784-1115 (A. Ferrara); 0000-0002-3306-4233 (S. Ghodsi); 0000-0002-3137-6890 (C. Mougan); 0000-0001-5238-4550 (P. Reyero); 0000-0002-3920-5017 (K. M. Scott); 0000-0001-8084-5297 (L. State); 0000-0001-6560-8947 (X. Zhao); 0000-0002-1917-6087 (S. Ruggieri)

leveraging from the results of the NoBIAS research project, we contribute to the ongoing policy advice and best practices discussion, focusing on the European context.

Fair-AI aims at designing methods for detecting, mitigating, and controlling biases in AI-supported decision-making [4, 5]. Given its focus on bias and fairness, fair-AI has coalesced multiple fields concerned with, among other research lines, the fairness of decision-making (e.g., [6, 7, 8]); bias as a cognitive, technical, and socio-technical phenomenon (e.g., [9, 10, 11, 12]); and designing ML systems for social good (e.g., [13, 14, 15]). The state-of-the-art has been developing mainly on the technical side, sometimes reducing fair-AI problems to a numeric optimization problem under some fairness metric [16, 17, 18]. This hegemonic view on fair-AI problems has been increasingly criticized within the own field (e.g., [19, 20, 21]), which, in turn, has expanded the state-of-the-art. Additionally, it is important to include as part of the state-of-the-art the regulatory frameworks being developed, in particular within the European Union (EU) – such as the GDPR [22] and the AI Act [23] – to enforce fair-AI goals.

It is challenging, especially for the novel researcher and practitioner interested in fair-AI within the EU, to have a comprehensive view of the state-of-the-art. Therefore, the objectives and, in turn, contributions of Álvarez et al. [1] are twofold:

- *First,* we provide an up-to-date entry-point to the state-of-the-art of the multidisciplinary research on bias and fairness in AI. We take a bird's-eye view of the methods and resources, with links to specialized surveys, and of the issues and challenges related to policies on bias and fairness in AI. Such an overview provides guidance for both new researchers and AI practitioners.

- *Second,* we contribute toward the objective of providing policy advice and best practices for dealing with bias and fairness in AI by leveraging from the results of the NoBIAS research project. We present and discuss topics that emerged during the execution of the research project, whose focus was on legal challenges in the context of the EU legislation, and on understanding, mitigating, and accounting for bias from a multidisciplinary perspective.

The NoBIAS project. The NoBIAS project (January 2020 - June 2024) was a Marie Skłodowska-Curie Innovative Training Network funded by the European Union's Horizon 2020 research and innovation program. The core objective of NoBIAS was to research and develop novel interdisciplinary methods for AI-based decision-making without bias. [1]

Figure 1 shows the project's architecture. The *Bias Management Layer* is made up of the various components contributed by the research projects of the 15 NoBIAS Early-Stage Researchers (ESRs). Together, these components aim to achieve three research objectives: understanding bias, mitigating bias, and accounting for bias in data and AI-systems. An orthogonal *Legal Layer* provides the necessary EU legal grounds. The purpose is not to produce one single bias management framework but rather to combine technologies and techniques for generating bias-aware AI-systems in different application domains and contexts.

Paper structure. Following the objectives, the paper is divided into two main sections.

In the ***Landscape of policies on bias and fairness in AI*** section, we provide a concise overview of the state-of-the-art for fair-AI methods and policy topics. In this section, we point

[1] For more information, visit: https://nobias-project.eu/

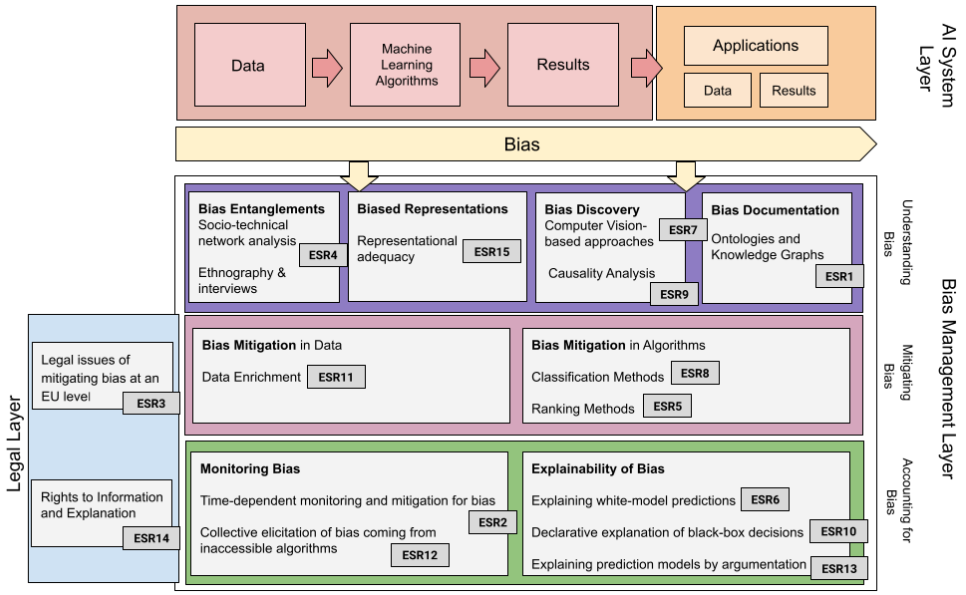

Figure 1: The NoBIAS architecture Each Early-Stage Researcher (ESR) focused on one aspect of the project's architecture, totaling 15 ESRs. Visit: https://nobias-project.eu/ for more information.

to the main contributions and resources in the area, providing guidance for both researchers and practitioners. First, we cover ***Fair-AI methods and resources***, in which we explore the fairness metrics (group-level, individual-level, and causality-based) [24, 25, 26, 27, 28, 29], tracing back their origins to fields like Philosophy and Economics [30, 6, 31, 32]. We also discuss common applications (e.g., computer vision [33] and recommender systems [34]) and popular standardization initiatives (e.g., the IEEE P7003™ Standard[2]). Second, we cover ***Policies on bias and fairness in AI***, in which we discuss policy and guidelines inventories (e.g., the OECD.AI Policy Observatory[3]); the option not to use AI (e.g., the Stop LAPD Spying Coalition[4]); documentation practices for bias (e.g., [35, 36, 37, 38]); and EU legal regimes and discussions around them (e.g., [23, 22, 39, 40, 41]); among other topics.

In the ***Lessons from the NoBIAS project*** section, we discuss policy advice and best practices resulting from the execution of the NoBIAS research project. Here, we take a critical view on the literature, focusing on findings from the NoBIAS project (e.g., [42, 43, 44, 45, 33, 46, 47, 48, 49]). We argue that the issues discussed are relevant, but not sufficiently developed or acknowledged in the fair-AI literature. Thus, this section further enriches the state-of-the-art. This section is organized according to the NoBIAS architecture in Figure 1. We cover the NoBIAS *Bias Management Layer* through the subsections ***Understanding bias***, ***Mitigating bias***, and ***Accounting for bias*** as well as the NoBIAS *Legal Layer* through the subsection ***Legal challenges of bias in AI***. Each subsection is further divided into relevant topics/themes. For

[2]https://standards.ieee.org/project/7003.html
[3]https://oecd.ai/en/dashboards/overview
[4]https://stoplapdspying.org/wp-content/uploads/2018/05/Before-the-Bullet-Hits-the-Body-May-8-2018.pdf

Legal Layer

- AI models often lack the auxiliary causal knowledge required to prove anti-discrimination cases as these require to show that the decision is *because of* the protected attribute.
- AI models' complexity and opaqueness make it difficult to identify individuals and groups that are treated unfairly.
- The design of AI models requires to agree on and to operationalise legal and ethical principles.
- Transparency and accountability of AI systems are a way to overcome the hegemonic theory of fairness, which reduces fairness problems to quantitative metric optimization.
- There are synergies and frictions in the EU legal framework between data protection law and non-discrimination law, which demand for an integrated and interdisciplinary techno-legal framework of bias management.

Figure 2: Legal Layer: challenges, policy advice, and best practices.

Bias Management Layer - Understanding Bias

- We should acknowledge that there are many forms of bias, with different roots and effects.
- The "ground-truth" is a myth. It does not exist in a structurally unjust and unequal society.
- Data curation in AI should import source criticism and archival practices from historical and humanistic disciplines.
- There is an hyper-fixation on data as the primary source of bias, but the whole AI pipeline needs to be addressed, including the data annotation process and data labourers' exploitation.
- Different data types require specific regulatory guidelines and standards.

Figure 3: Bias Management Layer – Understanding Bias: challenges, policy advice, and best practices.

instance, in ***Understanding bias*** we discuss the subjectivity of bias; argue that the notion of ground-truth can be itself biased; provide source criticism and archival practices on bias documentation; discuss data annotation; and present data types as a source of bias.

Each of these NoBIAS subsections is summarized in the form of a set of challenges, policy advice, and best practices aimed at all stakeholders. We present two of these below – Figures 2 and 3 – as representative examples. Each item listed in Figure 2 corresponds to a fair-AI topic discussed within the subsection ***Legal challenges of bias in AI***. Similarly, each item in Figure 3 corresponds to a topic discussed within the subsection ***Understanding bias***. All items are substantiated using the relevant fair-AI literature. Naturally, the choice of topic (i.e., item) was conditioned by the works of the NoBIAS ESRs.

Conclusion. In this work we provide a comprehensive introduction to the multidisciplinary and growing fair-AI literature. Using the NoBIAS research project as a guide, we extend the current discussion around the state-of-the-art by focusing on themes studied throughout the project. Leveraging on the NoBIAS architecture (Figure 1), we dwell into ongoing fair-AI research topics,

position these topics within the EU regulatory framework, and provide best practices and policy advice to the general practitioner (e.g., Figure 2). While we do not claim for their completeness, we hope that the policy advice and best practices provided in this paper will contribute to the conventional wisdom in research of and the ongoing discussion on managing bias and fairness in AI. Please refer to Álvarez et al. [1] for a complete discussion.

Acknowledgments. This work has received funding from the European Union's Horizon 2020 research and innovation program under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "NoBIAS - Artificial Intelligence without Bias". This work reflects only the authors' views and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains. Jose M. Alvarez and Salvatore Ruggieri are also partially supported by the European Community H2020-EU.2.1.1 program under the G.A. 952215 *Tailor*.

References

[1] J. M. Álvarez, A. Bringas-Colmenarejo, A. Elobaid, S. Fabbrizzi, M. Fahimi, A. Ferrara, S. Ghodsi, C. Mougan, I. Papageorgiou, P. R. Lobo, M. Russo, K. M. Scott, L. State, X. Zhao, S. Ruggieri, Policy advice and best practices on bias and fairness in AI, Ethics Inf. Technol. 26 (2024) 31.

[2] R. Shelby, S. Rismani, K. Henne, A. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla-Akbari, J. Gallegos, A. Smart, E. Garcia, G. Virk, Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction, in: AIES, ACM, 2023, p. 723–741.

[3] K. Martin, Ethical implications and accountability of algorithms, Journal of Business Ethics 160 (2019) 835–850.

[4] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, P. Hall, Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, Technical Report 1270, NIST Special Publication, 2022.

[5] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernández, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, S. Staab, Bias in data-driven Artificial Intelligence systems - An introductory survey, WIREs Data Mining Knowl. Discov. 10 (2020).

[6] B. Hutchinson, M. Mitchell, 50 years of test (un)fairness: Lessons for machine learning, in: FAT, ACM, 2019, pp. 49–58.

[7] B. Friedman, H. Nissenbaum, Bias in computer systems, ACM Trans. Inf. Syst. 14 (1996) 330–347.

[8] S. Lowry, G. Macpherson, A blot on the profession, British Medical Journal 296 (1986) 657–658.

[9] M. G. Haselton, D. Nettle, P. W. Andrews, The evolution of cognitive bias, in: E. N. Zalta (Ed.), Handbook of Evolutionary Psychology, John Wiley & Sons Inc., 2005, p. 724–746.

[10] T. Hellström, V. Dignum, S. Bensch, Bias in machine learning - what is it good for?, in:

NeHuAI@ECAI, volume 2659 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 3–10.
[11] D. A. Grimes, K. F. Schulz, Bias and causal associations in observational research, Lancet 359 (2002) 248–252.
[12] D. Danks, A. J. London, Algorithmic bias in autonomous systems, in: IJCAI, ijcai.org, 2017, pp. 4691–4697.
[13] D. Pedreschi, S. Ruggieri, F. Turini, Discrimination-aware data mining, in: KDD, ACM, 2008, pp. 560–568.
[14] F. Kamiran, T. Calders, Classifying without discriminating, in: Int. Conference on Computer, Control and Communication, IEEE, 2009, pp. 1–6.
[15] F. Kamiran, A. Karim, X. Zhang, Decision theory for discrimination-aware classification, in: ICDM, IEEE Computer Society, 2012, pp. 924–929.
[16] S. Ruggieri, J. M. Álvarez, A. Pugnana, L. State, F. Turini, Can we trust fair-AI?, in: AAAI, AAAI Press, 2023, pp. 15421–15430.
[17] A. N. Carey, X. Wu, The statistical fairness field guide: Perspectives from social and formal sciences, AI Ethics 3 (2023) 1–23.
[18] L. Weinberg, Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ML fairness approaches, J. Artif. Intell. Res. 74 (2022) 75–109.
[19] K. Wagstaff, Machine learning that matters, in: ICML, icml.cc / Omnipress, 2012.
[20] B. D. Mittelstadt, S. Wachter, C. Russell, The unfairness of fair machine learning: Levelling down and strict egalitarianism by default, CoRR abs/2302.02404 (2023).
[21] T. Scantamburlo, Non-empirical problems in fair machine learning, Ethics Inf. Technol. 23 (2021) 703–712.
[22] European Parliament, Council of the European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal of the European Union L 119 (2016). URL: http://data.europa.eu/eli/reg/2016/679/oj.
[23] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying down harmonised rules on Artificial Intelligence (AI Act) and amending certain Union legislative acts, 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.
[24] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, A. C. Penco, A clarification of the nuances in the fairness metrics landscape, Scientific Reports 12 (2022) 4209.
[25] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (2021) 115:1–115:35.
[26] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, Fairness in criminal justice risk assessments: The state of the art, Sociological Methods & Research 50 (2021) 3–44.
[27] S. Verma, J. Rubin, Fairness definitions explained, in: FairWare@ICSE, ACM, 2018, pp. 1–7.
[28] I. Zliobaite, Measuring discrimination in algorithmic decision making, Data Min. Knowl. Discov. 31 (2017) 1060–1089.
[29] S. Caton, C. Haas, Fairness in machine learning: A survey, ACM Comput. Surv. (2024) to appear.

[30] M. S. A. Lee, L. Floridi, J. Singh, Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics, AI Ethics 1 (2021) 529–544.
[31] R. Binns, Fairness in machine learning: Lessons from political philosophy, in: FAT, volume 81 of *Proc. of Machine Learning Research*, PMLR, 2018, pp. 149–159.
[32] A. Romei, S. Ruggieri, A multidisciplinary survey on discrimination analysis, Knowl. Eng. Rev. 29 (2014) 582–638.
[33] S. Fabbrizzi, S. Papadopoulos, E. Ntoutsi, I. Kompatsiaris, A survey on bias in visual datasets, Comput. Vis. Image Underst. 223 (2022) 103552.
[34] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, X. He, Bias and debias in recommender system: A survey and future directions, ACM Trans. Inf. Syst. 41 (2023) 67:1–67:39.
[35] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. M. Wallach, H. D. III, K. Crawford, Datasheets for datasets, Commun. ACM 64 (2021) 86–92.
[36] I. D. Raji, J. Yang, ABOUT ML: annotation and benchmarking on understanding and transparency of machine learning lifecycles, CoRR abs/1912.06166 (2019).
[37] J. Stoyanovich, S. Abiteboul, B. Howe, H. V. Jagadish, S. Schelter, Responsible data management, Commun. ACM 65 (2022) 64–74.
[38] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, P. Barnes, Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing, in: FAT*, ACM, 2020, pp. 33–44.
[39] I. Mendoza, L. A. Bygrave, The right not to be subject to automated decisions based on profiling, EU Internet Law: Regulation and Enforcement (2017) 77–98.
[40] Article 29 Data Protection Working Party, Guidelines on automated individual decision-making and profiling for the purposes of regulation 2016/679 (wp251rev.01), 2018. URL: https://ec.europa.eu/newsroom/article29/items/612053.
[41] A. Balayn, S. Gürses, Beyond debiasing: Regulating AI and Its Inequalities, Technical Report, European Digital Rights (EDRi), 2021.
[42] A. Ferrara, L. E. Noboa, F. Karimi, C. Wagner, Link recommendations: Their impact on network structure and minorities, in: WebSci, ACM, 2022, pp. 228–238.
[43] L. State, M. Fahimi, Careful explanations: A feminist perspective on XAI, in: EWAF, volume 3442 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.
[44] L. State, H. Salat, S. Rubrichi, Z. Smoreda, Explainability in practice: Estimating electrification rates from mobile phone data in senegal, CoRR abs/2211.06277 (2022).
[45] P. R. Lobo, E. Daga, H. Alani, M. Fernández, Semantic web technologies and bias in Artificial Intelligence: A systematic literature review, Semantic Web 14 (2023) 745–770.
[46] C. Mougan, J. M. Álvarez, S. Ruggieri, S. Staab, Fairness implications of encoding protected categorical attributes, in: AIES, ACM, 2023, pp. 454–465.
[47] K. M. Scott, S. M. Wang, M. Miceli, P. Delobelle, K. Sztandar-Sztanderska, B. Berendt, Algorithmic tools in public employment services: Towards a jobseeker-centric perspective, in: FAccT, ACM, 2022, pp. 2138–2148.
[48] J. M. Álvarez, K. M. Scott, B. Berendt, S. Ruggieri, Domain adaptive decision trees: Implications for accuracy and fairness, in: FAccT, ACM, 2023, pp. 423–433.
[49] J. M. Álvarez, S. Ruggieri, Counterfactual situation testing: Uncovering discrimination under fairness given the difference, in: EAAMO, ACM, 2023, pp. 2:1–2:11.