Trustworthy AI at KDD Lab

Fosca Giannotti[3], Riccardo Guidotti[1], Anna Monreale[1], Luca Pappalardo[2], Dino Pedreschi[1], Roberto Pellungrini[3], Francesca Pratesi[2,*], Salvatore Rinzivillo[2], Salvatore Ruggieri[1], Mattia Setzu[1] and Rosaria Deluca[2]

[1]*Computer Science Department, Largo Pontecorvo 3, Pisa, 56127, Italy*
[2]*ISTI - CNR, via Moruzzi 1, Pisa, 56127, Italy*
[3]*Scuola Normale Superiore, Piazza dei Cavalieri 7, Pisa, 56126, Italy*

Abstract
This document summarizes the activities regarding the development of Responsible AI (Responsible Artificial Intelligence) conducted by the Knowledge Discovery and Data mining group (KDD-Lab), a joint research group of the Institute of Information Science and Technologies "Alessandro Faedo" (ISTI) of the National Research Council of Italy (CNR), the Department of Computer Science of the University of Pisa, and the Scuola Normale Superiore of Pisa.

Keywords
Fairness, Privacy, Explainability, Trustworthy AI, Social AI

1. Introduction

Big Data typically describes different dimensions of the daily social life and are the heart of a knowledge society, where the understanding of social phenomena is sustained by the knowledge extracted from the miners of big data across the various social dimensions by using data mining, machine learning and AI technologies. The worrying side of the story is that this data also describes in detail personal or even sensitive aspects of our lives, thus, privacy leaks, security threats or discriminatory events can occur when Big Data are processed. We are particularly aware of ethical issues related to the processing of personal data, and we were precursors in the ethical management of personal data, especially in the fields of privacy, fairness, and explainabilty.

1.1. Lab Unit

The activity summarized in this paper is pursued by the Knowledge Discovery and Data mining group (KDD-Lab), a joint research group of the ISTI-CNR, the University of Pisa, and the Scuola Normale Superiore.

1.2. Projects

We are actively participating in several projects, having ethical issues and solutions in AI as central topic:

- XAI - Science and technology for the eXplanation of AI decision making[1] (ERC-AdG, GA 834756);
- SoBigData++ - European Integrated Infrastructure for Social Mining and Big Data Analytics[2] (H2020 RI, GA 871042);
- TAILOR - Foundations of Trustworthy AI - Integrating Reasoning, Learning and Optimization[3] (H2020 NoE, GA 952215);
- HumanE-AI-Net - HumanE AI Network[4] (H2020 NoE, GA 761758);
- LeADS - Legality Attentive Data Scientists[5] (MCSA, GA 956562);
- NoBIAS - Artificial Intelligence without Bias[6] (MCSA, GA 860630);
- FINDHR - Fairness and Intersectional Non-Discrimination in Human ecommendation[7] (HE, GA 101070212);
- CREXDATA - Critical Action Planning over Extreme-Scale Data (GA 101092749);
- FAIR - Future Artificial Intelligence Research[8] (Next Gen EU);
- SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics (Next Gen EU);

[1]https://xai-project.eu/
[2]https://plusplus.sobigdata.eu/
[3]https://tailor-network.eu/
[4]https://www.humane-ai.eu/
[5]https://www.lider-lab.it/test/leads/
[6]https://nobias-project.eu/
[7]https://https://findhr.eu//
[8]https://future-ai-research.it

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
*Corresponding author.
fosca.giannotti@sns.it (F. Giannotti); riccardo.guidotti@unipi.it (R. Guidotti); anna.monreale@unipi.it (A. Monreale); luca.pappalardo@isti.cnr.it (L. Pappalardo); pedre@di.unipi.it (D. Pedreschi); roberto.pellungrini@sns.it (R. Pellungrini); francesca.pratesi@isti.cnr.it (F. Pratesi); salvatore.rinzivillo@isti.cnr.it (S. Rinzivillo); salvatore.ruggieri@unipi.it (S. Ruggieri); mattia.setzu@di.unipi.it (M. Setzu); rosaria.deluca@isti.cnr.it (R. Deluca)

- TANGO - It takes two to tango: a synergistic approach to human-machine decision making (HE).

1.3. Organization of the paper

In the next section, we will describe, for the scientific themes related to *Trustworthy AI*, the work that has been carried out from all the people of the KDD Lab[9], with particular focus on the EU projects listed in Section 1.2.

2. Research Activities

Big Data analytics and *AI* are not necessarily enemies of *Ethics*. Sometimes many practical and impactful services based on Big Data analytics and machine learning (ML) can be designed in such a way that the quality of results can coexist with fairness, privacy protection, and transparency. The overall scientific objective regarding the Responsible AI is to develop the scientific foundations for *Trustworthy AI*. Trustworthy AI is high on both the political and scientific agenda for AI [1], and it is considered one of the hallmarks of AI made in Europe.

We especially focus on four ethical dimensions (Explainability, Fairness, Privacy, and Trustworthiness), and on the intertwining aspects that bound these dimensions. In the following, we will briefly outline, for each dimension, our main goals, the ongoing activities, and the open challenges that we are going to face in the projects listed in Section 1.2. However, for all the activities regarding these topics, we also: *(1)* engage the scientific community, e.g., organizing inclusive events in different conference such as international workshops [10] or special issues on leading journals[11], to foster collaborations and the cross-contamination; *(2)* develop focused solutions to target research problems with links to models and formalisms studied by the foundational themes; *(3)* generalize these solutions into reusable AI techniques and guidelines for standardized processes.

2.1. Explainable AI for decision making

The impressive performance of AI systems in prediction, recommendation, and decision-making support is generally reached by adopting complex ML models that often "hide" the logic of their internal processes. As a consequence, such models are often referred to as "black-box" models. AI-based systems are likely to lead to decisions that we do not fully understand, and, even worse, decisions that are likely to violate ethical principles. In 2018, the European Parliament introduced in the GDPR[12] a set of clauses for automated decision-making in terms of a right of explanation for all individuals to obtain "meaningful explanations of the logic involved" when automated decision-making takes place. Also, in 2019, the High-Level Expert Group on AI presented the ethics guidelines for trustworthy AI [1].

Despite divergent opinions among legals regarding these clauses, everybody agrees that the need for the implementation of such a principle is urgent and that it is a huge open scientific challenge. As a reaction to these practical and theoretical ethical issues, in the last years, we have witnessed the rise of a plethora of explanation methods for black-box models [2, 3] both from academia and from industries. eXplainable Artificial Intelligence (XAI) emerged as a research field to investigate methods to create or complement AI models whose internal logic is accessible and interpretable, thus making the decision-making process human-understandable, and helping people make better decisions preserving (and expanding) human autonomy. For this aim, we defined different research lines and we have been achieving the following results:

- We designed algorithms for the inference of local explanations for revealing the decision rationale for a specific case, developing novel algorithms such as: *(a)* an algorithm for local explanation that learns factual and counterfactual logic rules [4, 5]; *(b)* a set of methods that moves the generation of explanations into a latent space to produce exemplars and counter-exemplars for images [6, 7], for time-series [8, 9], and for text [10]; *(c)* approaches tailored to explaining decision of text classifiers. In particular, we designed an explainer that produces an explanation of a document classifier by generating new documents in its vicinity through word replacement, either by replacing words in the document with their synonyms, antonyms, hyponyms, hypernyms, and definitions. The approach preserves the structure of the original text as it generates synthetic text samples through the use of an ontology [11]. Moreover, we also developed an explainability algorithm for Transformer-based models fine-tuned on Natural Language Inference, Semantic Text Similarity, or Text Classification tasks [12]. The explanation is obtained by extracting a set of facts from the input data, subsuming it by abstraction, and generating a set of weighted triples as explanation; *(d)* we studied and advanced the state-of-the-art for counterfactual explanations

[9] https://kdd.isti.cnr.it/people

[10] https://kdd.isti.cnr.it/xkdd2022/, http://iail2022.isti.cnr.it/

[11] Special Issue on Ethics, Law and Responsible AI at journal Ethics and Information Technology, https://www.springer.com/journal/10676/, Special Issue on Trustworhty AI at ACM Computing Surveys https://dl.acm.org/pb-assets/static_journal_pages/csur/pdf/CSUR-CFP-Trustworthy-AI_083022-1661888117770.pdf

[12] https://ec.europa.eu/justice/smedataprotect/

both by design [13, 14] and post-hoc [15, 16].

- We explored languages for expressing explanations, in terms of both expressive logic rules (with statistical and causal interpretation) and models that capture the detailed data generation behind specific deep learning models. We designed a framework that composes rules representing local explanations into a global explanation by merging theories, a form of logical meta-reasoning [17].
- We explored the opportunity of creating a co-design methodology to develop a human-centered, explainable AI system for decision support [18]. Specifically, we designed a prototype-test-redesign loop involving healthcare providers as end-users, which we then used to refine an explanation algorithm and its user interface. We first presented the XAI technique's conception based on the patients data and healthcare application requirements. Then, we developed the initial prototype of the explanation user interface, and perform a user study to test its perceived trustworthiness and collect healthcare providers' feedback. We finally exploit the users' feedback to co-design a more human-centered XAI user interface taking into account cognitive design principles such as progressive disclosure of information.
- We investigated unexplored aspects such as the explanation of complex models like Siamese Networks used in few-shot and zero -shot learning [19], and the exploitation of causal discovery to improve the explainers effectiveness [20].
- We endorsed the creation of a common ground for researchers working on explanation from different domains, we developed a platform consisting of two parts: *(i)* a software library that integrates a wide set of explanation methods; *(ii)* a dedicated visual interface to let the user to interact with the explanation. We survey Explanation methods focusing on benchmarking [21].

In this field, we aim to push forward the research in both directions: *(a)* developing post-hoc explanations that given a black-box model aims to reconstruct its logic, and *(b)* drive towards explainable-by-design models. This must be done having the performance and interpretability trade-off in mind, trying to achieve both by following a human-centered methodology to produce explanations suitable to the cognitive skills of their users.
Finally, we need to explore what is still missing: for example, a formalism for explanations, and standards and metrics to quantify the grade of comprehensibility of an explanation for humans. These standards need to take into account the research results from the HCI, DataVis, and Cognitive Sciences communities.

2.2. Respect for privacy

The main goal of the scientific research on privacy is to design privacy-preserving solutions that guarantee to achieve both privacy protection (i.e., not revealing any personal or sensitive information about individuals or companies whose data are referring to) and utility of the data-driven services. We have been achieving interesting results in the following research lines:

- *Privacy-by-design paradigm.* We explored the potentiality of privacy-by-design paradigm in designing and developing technological frameworks to counter the threats of undesirable effects of privacy violation, without obstructing the knowledge discovery opportunities of big data analytical technologies, by inscribing privacy protection into the data processing by design. We applied this principle in different applications, such as mobility data analytics [22] or call activities [23].
- *Privacy Risk Assessment.* We developed methodologies for systematically evaluating the risk of re-identification of all the individuals in a certain dataset [24]. This framework can be applied when it is not totally clear what kind of information is owned by a malicious third party, and it has been tested in different settings with different kinds of data, such as mobility data [24], retail data [25] and psychometric profiles [26]. We also proposed an adversarial model based on this framework and developed an optimization algorithm tailored for human mobility data to determine the most damaging adversarial behavior w.r.t. the privacy of the individuals in the data [27]. We also studied the privacy risk of federated leaning systems and we defined a new approach aiming to reduce by generalization the assessed risks [28].
- *Privacy Risk Prediction.* We extended the previous framework, allowing to obtain a prediction of the privacy risk for previously unseen individuals, i.e., not belonging to the starting dataset [29].

Even though privacy is one of the first human rights that has been considered in legal frameworks, and a lot of work has been done in the scientific literature, there is still the need to investigate new methodologies and approaches for: *(a)* defining formally and detecting automatically privacy risks raised by AI systems handling different kinds of personal data; *(b)* designing data anonymization algorithms that are robust to sophisticated attacks; *(c)* designing AI algorithms that respect by-design privacy constraints also in distributed scenarios, where individuals can cooperate for a common learning goal but with different privacy requirements; *(d)* investigating existing measures (or creating new ones) to evaluate the privacy risk of novel or unusual kinds of data, especially with respect to the interplay with other ethical dimensions.

2.3. Fairness, equity and justice by-design

AI models' outputs can be biased against specific individuals or groups [30]. The most relevant effect of such a bias is unfairness or even illegal discrimination against protected-by-law social groups [31]. Equity requires that people are treated according to their needs, which does not mean all people are treated equally [32]. Justice is the "fair and equitable treatment of all individuals under the law" [33]. Fair AI models are designed to prevent biased decisions in algorithmic decision making. Quantitative definitions of fairness have been introduced in philosophy, economics, and machine learning in the last 50 years [34], with more than 20 different definitions of fairness appeared thus far in the computer science literature [35]. We contributed to the definition of:

- *Group fairness* metrics, measuring the statistical difference in distributions of decisions across social groups, with the pioneering works [36].
- *Individual fairness* metrics, binding the distance among the decision space and the feature space describing people's characteristics [37].
- *Causal fairness* metrics, which exploit knowledge beyond observational data to infer causal relations between features and decisions, and to estimate interventional consequences [38].

Based on these metrics, methods and tools have been proposed for:

- Bias detection (*discrimination discovery* or *fairness testing*), including our approaches [36, 39, 40] and case studies [41, 42].
- Dataset de-biasing and data processing (*preprocessing approaches*), including our work connecting fairness and privacy [43].
- Training or correcting AI models and representations through fair algorithms (*in-processing* and *post-processing approaches*), including [44].
- *Monitoring* models' decisions [45, 46].

As with other quality objectives, the choice of a fairness metric is crucial for optimizing and for auditing AI models [47]. We have recently supported the critiques to the hegemonic theory of fairness [48], which reduces the problem to a numeric optimization of some metrics [49]. Pathways for research include, in our view, multistakeholders participatory design, integration with other trustworthy tools for AI, notably explanation methods, and the option to reject unfair AI outcomes.

2.4. Trustworthy AI as a whole

We studied potential tensions but also synergies among these ethical dimensions. We have been exploring the bounds between explainability and fairness [50] and between privacy and explainability [51, 52]. Concerning the latter one, we studied both: *i)* how XAI techniques help individuals to acquire awareness on their potential privacy risks providing them with insights on which behavior contributes most [51, 53]; *ii)* how transparency may jeopardize the privacy risks of individuals represented in data used for training ML models [52].

We also developed an ethico-legal framework for responsible data science [54] and we promoted general aspects such as digital ecosystem of trust [55].

In addition, a goal that we would like to pursue is also to build Reference Datasets, which may boost both research along the dimensions above and their combinations, and may allow to compare the various proposed solutions among them acting as benchmarks, together with evaluation criteria, to assess whether proposed AI systems are Trustworthy or not.

2.5. The social dimension of AI

The rise of socio-technical systems (STS) in which humans interact with various forms of AI systems, including assistants and recommenders (AIs), amplifies the possibility for the emergence of large-scale social behaviour, possibly with unintended negative consequences. While AIs may generate individually "good" suggestions, the sum of many suggestions can have unintended outcomes because users' choices, influenced by these suggestions, interfere with each other on top of shared resources. For example, GPS navigation systems suggest directions that make sense from an individual perspective but may create chaos if too many drivers are directed on the same route; and personalised recommendations on social media often make sense to the user but may artificially amplify echo chambers, filter bubbles, and radicalisation. This happens because AIs are based on ML models, generating a feedback loop: users' preferences determine the training datasets on which AIs are trained; the trained AIs then exert a new influence on users' subsequent preferences, influencing the next round of training, and so on.

As an example, we conducted a study in Milan to investigate the impact of navigation systems on the urban environment in terms of CO2 emissions [56]. We simulated the behavior of vehicles by assuming that they would either follow a commercial navigation system or deviate randomly from the fastest route. The results showed that blindly following the recommendations of a navigation app can lead to traffic congestion in some areas of the city, resulting in increased travel time and global emissions. We also observed that adding controlled randomness to suggested routes results in a better distribution of traffic on the road network, leading to decreased travel time and emissions, without penalizing individual travellers. The study suggests the need for developing routing algo-

rithms that prioritize route diversification, as conforming to a limited set of routes can diminish diversity in drivers' behavior and lead to inefficient road network use.

Understanding the impact of AIs on STS is emerging as another challenging dimension of trustworthy AI, which may enable the unprecedented opportunity to intervene on such STS to proactively help achieve important agreed goals with a better balance of individual and collective interests. However, achieving such a broader understanding requires a change of perspective that embraces complexity science and the trans-disciplinary integration of network science, AI, and computational social science.

Acknowledgments

This work is supported by the EU – Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities", G.A. n.871042 "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics", by the PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme and by the EU – NextGenerationEU – National Recovery and Resilience Plan – Project: "SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics" – Prot. IR0000013 – Avviso n. 3264 del 28/12/2021.

References

[1] High Level Expert Group, Ethics guidelines for trustworthy ai, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[2] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018).

[3] A. B. Arrieta, et al., Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020).

[4] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, F. Turini, Factual and counterfactual explanations for black box decision making, IEEE Intelligent Systems 34 (2019).

[5] R. Guidotti, A. Monreale, S. Ruggieri, F. Naretto, F. Turini, D. Pedreschi, F. Giannotti, Stable and actionable explanations of black-box models through factual and counterfactual rules, Data Mining and Knowledge Discovery (2022).

[6] R. Guidotti, A. Monreale, D. Matwin, Stan & Pedreschi, Explaining image classifiers generating exemplars and counter-exemplars from latent representations, in: AAAI, volume 34, 2020.

[7] C. Metta, R. Guidotti, Y. Yin, P. Gallinari, S. Rinzivillo, Exemplars and counterexemplars explanations for skin lesion classifiers, in: HHAI2022: Augmenting Human Intellect, IOS Press, 2022.

[8] R. Guidotti, A. Monreale, F. Spinnato, D. Pedreschi, F. Giannotti, Explaining any time series classifier, in: CogMI, IEEE, 2020.

[9] F. Spinnato, R. Guidotti, M. Nanni, D. Maccagnola, G. Paciello, A. B. Farina, Explaining crash predictions on multivariate time series data, in: DS, 2022.

[10] O. Lampridis, L. State, R. Guidotti, S. Ruggieri, Explaining short text classification with diverse synthetic exemplars and counter-exemplars, Machine Learning (2022).

[11] M. Sarvmaili, R. Guidotti, A. Monreale, A. Soares, Z. Sadeghi, F. Giannotti, D. Pedreschi, S. Matwin, A modularized framework for explaining black box classifiers for text data, in: AI, Canadian AI Association, 2022.

[12] M. Setzu, A. Monreale, P. Minervini, Triplex: Triple extraction for explanation, in: IEEE Conference on Cognitive Machine Intelligence, 2021.

[13] F. Bodria, R. Guidotti, F. Giannotti, D. Pedreschi, Transparent latent space counterfactual explanations for tabular data, in: DSAA 2022, IEEE, 2022.

[14] F. Bodria, R. Guidotti, F. Giannotti, D. Pedreschi, Interpretable latent space to enable counterfactual explanations, in: Discovery Science, Springer, 2022.

[15] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery (2022).

[16] R. Guidotti, S. Ruggieri, Ensemble of counterfactual explainers, in: Discovery Science, Springer, 2021.

[17] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti, Glocalx - from local to global explanations of black box AI models, A.I. 294 (2021).

[18] C. Panigutti, A. Beretta, D. Fadda, F. Giannotti, D. Pedreschi, A. Perotti, S. Rinzivillo, Co-design of human-centered, explainable ai for clinical decision support, ACM Trans. Interact. Intell. Syst. (2023).

[19] A. Fedele, R. Guidotti, D. Pedreschi, Explaining siamese networks in few-shot learning for audio data, in: DS, 2022.

[20] M. Cinquini, R. Guidotti, Calime: Causality-aware local interpretable model-agnostic explanations, arXiv preprint arXiv:2212.05256 (2022).

[21] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, Benchmarking and survey of explanation methods for black box models, volume abs/2102.13076, 2021.

[22] A. Monreale, S. Rinzivillo, F. Pratesi, F. Giannotti,

D. Pedreschi, Privacy-by-design in big data analytics and social mining, EPJ Data Science 10 (2014).
[23] F. Pratesi, L. Gabrielli, P. Cintia, A. Monreale, F. Giannotti, Primule: Privacy risk mitigation for user profiles, Data & Knowledge Engineering 125 (2020).
[24] F. Pratesi, A. Monreale, R. Trasarti, F. Giannotti, D. Pedreschi, T. Yanagihara, PRUDEnce: a system for assessing privacy risk vs utility in data sharing ecosystems, Transactions on Data Privacy 11 (2018).
[25] R. Pellungrini, A. Monreale, R. Guidotti, Privacy risk for individual basket patterns, in: ECML PKDD Workshops, 2018.
[26] G. Mariani, A. Monreale, F. Naretto, Privacy risk assessment of individual psychometric profiles, in: DS, 2021.
[27] R. Pellungrini, L. Pappalardo, F. Simini, A. Monreale, Modeling adversarial behavior against mobility data privacy, IEEE TITS 23 (2022).
[28] M. Fontana, F. Naretto, A. Monreale, A new approach for cross-silo federated learning and its privacy risks, in: PST, IEEE, 2021.
[29] R. Pellungrini, L. Pappalardo, F. Pratesi, A. Monreale, A data mining approach to assess privacy risk in human mobility data, ACM TIST 9 (2017).
[30] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, et al., Bias in data-driven artificial intelligence systems — an introductory survey, WIREs Data Mining and Knowledge Discovery 10 (2020).
[31] A. Altman, Discrimination, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, 2020.
[32] M. Minow, Equality vs. equity, American Journal of Law and Equality 1 (2021).
[33] J. Lehman, S. Phelps, et al., West's encyclopedia of American law, Thomson/Gale, 2004.
[34] B. Hutchinson, M. Mitchell, 50 years of test (un)fairness: Lessons for machine learning, in: FAT, ACM, 2019.
[35] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (2021).
[36] S. Ruggieri, D. Pedreschi, F. Turini, Data mining for discrimination discovery, ACM Trans. Knowl. Discov. Data 4 (2010).
[37] B. T. Luong, S. Ruggieri, F. Turini, k-nn as an implementation of situation testing for discrimination discovery and prevention, in: KDD, ACM, 2011.
[38] B. Qureshi, F. Kamiran, A. Karim, S. Ruggieri, D. Pedreschi, Causal inference for social discrimination reasoning, J. Intell. Inf. Syst. 54 (2020).
[39] S. Ruggieri, S. Hajian, F. Kamiran, X. Zhang, Anti-discrimination analysis using privacy attack strategies, in: ECML/PKDD, 2014.
[40] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, F. Giannotti, Discrimination- and privacy-aware patterns, Data Min. Knowl. Discov. 29 (2015).
[41] M. M. Manerba, R. Guidotti, L. C. Passaro, S. Ruggieri, Bias discovery within human raters: A case study of the jigsaw dataset, in: NLPerspectives@LREC, 2022.
[42] A. Romei, S. Ruggieri, F. Turini, Discrimination discovery in scientific project evaluation: A case study, Expert Syst. Appl. 40 (2013).
[43] S. Ruggieri, Using t-closeness anonymity to control for non-discrimination, Trans. Data Priv. 7 (2014).
[44] D. Pedreschi, S. Ruggieri, F. Turini, Measuring discrimination in socially-sensitive decision records, in: SDM, SIAM, 2009.
[45] K. Kenthapadi, H. Lakkaraju, P. Natarajan, M. Sameki, Model monitoring in practice: Lessons learned and open challenges, in: KDD, ACM, 2022.
[46] M. Fontana, F. Naretto, A. Monreale, F. Giannotti, Monitoring fairness in HOLDA, in: HHAI, volume 354 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2022, pp. 246–248.
[47] D. Pedreschi, S. Ruggieri, F. Turini, A study of top-k measures for discrimination discovery, in: SAC, ACM, 2012.
[48] L. Weinberg, Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ML fairness approaches, J. Artif. Intell. Res. 74 (2022).
[49] S. Ruggieri, J. M. Alvarez, A. Pugnana, L. State, F. Turini., Can we trust fair-AI?, in: AAAI Conference on Artificial Intelligence, 2023.
[50] C. Panigutti, A. Perotti, A. Panisson, P. Bajardi, D. Pedreschi, Fairlens: Auditing black-box clinical decision support systems, Information Processing & Management 58 (2021).
[51] F. Naretto, R. Pellungrini, A. Monreale, F. M. Nardini, M. Musolesi, Predicting and explaining privacy risk exposure in mobility data, in: DS, 2020.
[52] F. Naretto, A. Monreale, F. Giannotti, Privacy risk of global explainers, in: HHAI, volume 354 of *Frontiers in Artificial Intelligence and Applications*, 2022.
[53] F. Naretto, R. Pellungrini, F. M. Nardini, F. Giannotti, Prediction and explanation of privacy risk on mobility data with neural networks, in: PKDD/ECML Workshops, volume 1323, Springer, 2020.
[54] N. Forgó, S. Hänold, J. van den Hoven, T. Krügel, I. Lishchuk, R. Mahieu, A. Monreale, D. Pedreschi, F. Pratesi, D. van Putten, An ethico-legal framework for social data science, JDSA (2020).
[55] J. van den Hoven, G. Comandé, S. Ruggieri, J. Domingo-Ferrer, F. Musiani, F. Giannotti, F. Pratesi, M. Stauch, I. Lishchuk, Towards a digital ecosystem of trust: Ethical, legal and societal implications, Opinio Juris in Comparatione (2021).
[56] G. Cornacchia, M. Böhm, G. Mauro, M. Nanni, D. Pedreschi, L. Pappalardo, How routing strategies impact urban emissions, 2022.