Responsible AI at KDD Lab

Fosca Giannotti, Anna Monreale, Dino Pedreschi, Francesca Pratesi
Scuola Normale Superiore di Pisa & Istituto di Scienza e Tecnologia del Consiglio Nazionale delle Ricerche, Università di Pisa, Università di Pisa, Istituto di Scienza e Tecnologia del Consiglio Nazionale delle Ricerche
fosca.giannotti@isti.cnr.it, anna.monreale@unipi.it, dino.pedreschi@unipi.it, francesca.pratesi@isti.cnr.it

Abstract

This document summarizes the activities regarding the development of Responsible AI 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) and the Department of Computer Science of the University of Pisa.

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 complex 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 some personal or even sensitive aspects of our lives, thus, privacy leaks, security threats or discriminatory events can occur when Big Data are collected and 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 [Andrienko *et al.*, 2009], fairness [Hajian *et al.*, 2012], and explainabilty [Guidotti *et al.*, 2018].

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 Institute of Information Science and Technologies "Alessandro Faedo" (ISTI) of the National Research Council of Italy (CNR) and the Department of Computer Science of the University of Pisa.

1.2 Researchers involved

The list of researchers of our KDD-Lab group (both Junior and Senior, in alphabetical order) who are contributing in the Responsible AI topic is composed of: Jose Manuel Alvarez, Andrea Beretta, Isacco Beretta, Francesco Bodria, Giovanni Camarda, Eleonora Cappuccio, Martina Cinquini, Daniele Fadda, Michele Fontana, Fosca Giannotti, Riccardo Guidotti, Marta Marchiori Manerba, Carlo Metta, Anna Monreale, Francesca Naretto, Cecilia Panigutti, Dino Pedreschi, Roberto Pellungrini, Francesca Pratesi, Andrea Pugnana, Salvatore Rinzivillo, Salvatore Ruggieri, Mattia Setzu, Francesco Spinnato, Laura State, and Franco Turini.

1.3 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], funded by the European Union's Horizon 2020 Excellent Science European Research Council (ERC) programme under grant agreement No. 834756;
- **SoBigData++ - European Integrated Infrastructure for Social Mining and Big Data Analytics**[2], funded by the European Union's Horizon 2020 research and innovation programme under grant agreements No. 871042;
- **TAILOR - Foundations of Trustworthy AI - Integrating Reasoning, Learning and Optimization**[3], funded by the European Union's Horizon 2020 programme under grant agreement No. 952215;
- **HumanE-AI-Net - HumanE AI Network**[4], funded by the European Union's Horizon 2020 programme under grant agreement No. 761758;
- **SAI - Social explainable Artificial Intelligence**[5], under the EC CHIST-ERA program;
- **LeADS - Legality Attentive Data Scientists**[6], funded by the European Union's Marie Skłodowska-Curie programme under grant agreement No. 956562;
- **NoBIAS - Artificial Intelligence without Bias**[7], funded by European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860630.

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

1.4 Organization of the paper

In the following Section, we will describe, for each main scientific themes we are pursuing related to *responsible AI*, the work that has been carried out from all the people listed in Section 1.2 in the recent years, with particular focus on the European projects where we are involved (Section 1.3), and summarize the open challenges and possible future work.

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 KDD Lab has been pursuing these objectives since early 2000's, working on the goal to develop technological frameworks for countering the threats of undesirable, unlawful effects, which might derive from the use of data-driven technological systems.

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 (cf. the report of the High-Level Expert Group on AI [High Level Expert Group, 2019]), and it is considered one of the hallmarks of AI made in Europe. Developing trustworthy AI requires all of artificial intelligence to work in sync. The technological foundations for AI are learning, reasoning and optimisation, but they do not work seamlessly together, as they were developed independently. Therefore, realizing trustworthy AI is not possible without tightly integrating learning, reasoning and optimisation.

We especially focus on four ethical dimensions (Explainability, Fairness, Accountability, and Privacy), 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 1.3. However, for all the activities regarding these topic, we also: *(1)* engage the scientific community, e.g., organizing inclusive events[8] or special issues on leading journals[9], to foster collaborations and the cross-contamination; *(2)* develop focused solutions to two 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 Respect for privacy

The main goal of the scientific research on privacy are explore novel privacy-preserving solutions that guarantee to achieve both privacy (i.e., not revealing any personal or sensitive information about individuals or companies whose data are referring to) and utility of the implemented service.
For this aim, we:

- Have been explored the potentiality of privacy-by-design paradigm in designing and developing technological frameworks to counter the threats of undesirable, unlawful effects of privacy violation, without obstructing the knowledge discovery opportunities of social mining and big data analytical technologies, by inscribing privacy protection into the knowledge discovery process by design. We applied this principle in different applications, such as mobility data analytics [Monreale *et al.*, 2014] or call activities [Pratesi *et al.*, 2020].
- Have been developed privacy risk assessment methodologies for systematically evaluating the risk of re-identification of all the individuals in a certain dataset [Pratesi *et al.*, 2018]. 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 [Pratesi *et al.*, 2018] and retail data [Pellungrini *et al.*, 2018].
- Have been extended the previous framework, allowing to obtain a prediction of the actual individual privacy risk for previously unseen individuals, i.e., not belonging to the starting dataset [Pellungrini *et al.*, 2017].

Beside 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; *(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 (see below).

2.2 Explainable AI for decision making

So far, the usage of black boxes in AI and ML processes implied the possibility of inadvertently making wrong decisions, for example due to a systematic bias in training data collection. The main goal of this research theme is to focus on the urgent open challenge of how to construct meaningful explanations of opaque AI/ML systems in the context of AI based decision making, aiming at empowering individual against undesired effects of automated decision making, implementing the "right of explanation", helping people make better decisions preserving (and expand) human autonomy.
For this aim, we defined different research lines and we have been achieving the following results:

- Defining 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 produces local explanation in the form of factual and counterfactual logic rules [Guidotti *et al.*, 2019]; *(b)* a method that moves the generation of explanations into a latent space to produce exemplars and counter-exemplars [Guidotti *et al.*, 2020].

[8] https://sites.google.com/view/tailorwp3/meetingsworkshops/open-wp3-meeting-02092021

[9] Special Issue on Ethics, Law and Responsible AI at journal Ethics and Information Technology, https://www.springer.com/journal/10676/

- Exploring 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 that represent local explanations into a global explanation by merging theories, a form of logical meta-reasoning [Setzu *et al.*, 2021].
- Investigating 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 [Bodria *et al.*, 2021].

In this field, we will aim to pushing forward the research (i.e., proposing new Explainability methods) in both directions: *(a)* developing post-hoc explanation that given an opaque ML model (black box) aims to reconstruct its logic, and *(b)* driving towards transparent-by-design models that are explainable on their own. This must be done having the accuracy and interpretability trade-off in mind, trying to achieve both results.

Finally, we need to explore what is still missing: for example, a formalism for explanations, but also standards and metrics to quantify the grade of comprehensibility of an explanation for humans. Those standards need to take into account the research results from the HCI, DataVis, and Cognitive Sciences communities.

2.3 Fairness, equity and justice by-design

ML models can amplify existing bias coded in data or introduce new forms of bias [Ntoutsi *et al.*, 2020], resulting in unfair decisions in legal or ethical sense. In particular, AI-based systems may produce decisions or have impacts that are discriminatory or unfair, and this is specially true when AI is deployed in complex socio-technical systems.

There are essentially two lines of research, in this field:

- One consists in auditing AI systems for discrimination/segregation discovery, by investigating how decisions vary among social groups that differ w.r.t. sensitive variables [Ntoutsi *et al.*, 2020]. We have been mostly focused on this first line, analyzing the problems of discrimination discovery [Luong *et al.*, 2016] and segregation discoveries [Baroni e Ruggieri, 2018], and also providing tools for detecting these kinds of bias [Baroni e Ruggieri, 2019].
- The other line of research embeds the fairness value in the design of AI/ML models (fairness-by-design).

However, mis-representation in the data and how to address it is still under-investigated in the scientific community. Auditing AI-based systems is essential to discover cases of discrimination and to understand the reasons behind them and possible consequences (e.g., segregation). Methods for auditing AI-based systems for discrimination discovery typically investigate how decisions vary between social groups that differ w.r.t. sensitive variables. The perils of correlation analysis (for example, understanding causal influences among variables as a fundamental tool for dealing with bias) have been pointed out only recently.

The objective of equity can be also achieved by embedding the fairness value in the design of such systems (fairness-by-design) and by upholding that value (justice). A systematic approach that investigates how to build AI systems that respect by design fairness constraints for a variety of tasks such as classification, recommendation, resource allocation or matching is missing.

2.4 Accountability:

Accountability regards the governance of the design, development, and deployment of algorithmic systems, which takes into consideration all stakeholders and interactions with socio-technical systems. Specifically, we need to introduce and study techniques for data collection and analysis and processing that:

- Acknowledge the systemic bias and discrimination that may be present in datasets and models.
- Formalize fairness objectives based on notions from the social sciences, law, and humanistic studies.
- Build socio-technical systems incorporating these insights to minimize harm on historically disadvantaged communities and empower them.
- Introduce methods for decision validation, correction and participation in co-designing algorithmic systems.

It is also needed to define scientific and methodological measures, quality standards and procedures to better model the development process of learning methods.

2.5 Trustworthy AI as a whole

Last but not least, we also studied potential tensions but also synergies among these ethical dimensions. In particular, accuracy is a goal which is usually in contrast with all the ethical dimensions, i.e., there exist accuracy vs. fairness, accuracy vs. explainability, and accuracy vs. privacy trade-offs. We have been exploring the bounds between privacy and explainability [Naretto *et al.*, 2020], but also between explaiability and fairness [Ruggieri *et al.*, 2020; Panigutti *et al.*, 2021]. Finally, we develop an ethico-legal framework for responsible data science [Forgó *et al.*, 2020] and we promote general aspects such as digital ecosystem of trust and data altruism [van den Hoven *et al.*, 2021].

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 (or partial combination), and may allow to compare the various proposed solutions among them. Reference datasets may also act as benchmarks, together with evaluation criteria, to assess whether proposed AI systems are Trustworthy or not.

Riferimenti bibliografici

[Andrienko *et al.*, 2009] Gennady Andrienko, Natalia Andrienko, Fosca Giannotti, Anna Monreale, e Dino Pedreschi. Movement data anonymity through generalization.

In *Proceedings of the 2nd SIGSPATIAL ACM GIS 2009 International Workshop on Security and Privacy in GIS and LBS*. ACM, ACM, 2009.

[Baroni e Ruggieri, 2018] Alessandro Baroni e Salvatore Ruggieri. Segregation discovery in a social network of companies. *Journal of Intelligent Information Systems*, 51, 2018.

[Baroni e Ruggieri, 2019] Alessandro Baroni e Salvatore Ruggieri. Scube: A tool for segregation discovery. In *22nd International Conference on Extending Database Technology (EDBT)*, 2019.

[Bodria *et al.*, 2021] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, e Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. (2021).

[Forgó *et al.*, 2020] Nikolaus Forgó, Stefanie Hänold, Jeroen van den Hoven, Tina Krügel, Iryna Lishchuk, René Mahieu, Anna Monreale, Dino Pedreschi, Francesca Pratesi, e David van Putten. An ethico-legal framework for social data science. *International Journal of Data Science and Analytics*, 2020.

[Guidotti *et al.*, 2018] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, e Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2018.

[Guidotti *et al.*, 2019] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, e Franco Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34(6), 2019.

[Guidotti *et al.*, 2020] Riccardo Guidotti, Anna Monreale, e Dino Matwin, Stan Pedreschi. Explaining image classifiers generating exemplars and counter-exemplars from latent representations. In *Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 09, pp. 13665-13668)*, 2020.

[Hajian *et al.*, 2012] Sara Hajian, Anna Monreale, Dino Pedreschi, Josep Domingo-Ferrer, e Fosca Giannotti. Injecting discrimination and privacy awareness into pattern discovery. In *12th IEEE International Conference on Data Mining Workshops, ICDM Workshops, Brussels, Belgium, December 10, 2012*, page 360–369, 2012.

[High Level Expert Group, 2019] High Level Expert Group. Ethics guidelines for trustworthy ai, 2019. Last Accessed on 14 January 2022.

[Luong *et al.*, 2016] Binh Thanh Luong, Salvatore Ruggieri, e Franco Turini. Classification rule mining supported by ontology for discrimination discovery. In *IEEE ICDM Int. Workshop on Privacy and Discrimination in Data Mining (PDDM)*, 2016.

[Monreale *et al.*, 2014] Anna Monreale, Salvatore Rinzivillo, Francesca Pratesi, Fosca Giannotti, e Dino Pedreschi. Privacy-by-design in big data analytics and social mining. *European Physical Journal Data Science*, 10, 2014.

[Naretto *et al.*, 2020] Francesca Naretto, Roberto Pellungrini, Anna Monreale, Franco Maria Nardini, e Mirco Musolesi. Predicting and explaining privacy risk exposure in mobility data. In *International Conference on Discovery Science*, 2020.

[Ntoutsi *et al.*, 2020] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, e Steffen Staab. Bias in data-driven artificial intelligence systems — an introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10, 2020.

[Panigutti *et al.*, 2021] Cecilia Panigutti, Alan Perotti, André Panisson, Paolo Bajardi, e Dino Pedreschi. Fairlens: Auditing black-box clinical decision support systems. *Information Processing & Management*, 58(5), 2021.

[Pellungrini *et al.*, 2017] Roberto Pellungrini, Luca Pappalardo, Francesca Pratesi, e Anna Monreale. A data mining approach to assess privacy risk in human mobility data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9, 2017.

[Pellungrini *et al.*, 2018] Roberto Pellungrini, Anna Monreale, e Riccardo Guidotti. Privacy risk for individual basket patterns. In *ECML PKDD Workshops*, 2018.

[Pratesi *et al.*, 2018] Francesca Pratesi, Anna Monreale, Roberto Trasarti, Fosca Giannotti, Dino Pedreschi, e Tadashi Yanagihara. Prudence: a system for assessing privacy risk vs utility in data sharing ecosystems. *Transactions on Data Privacy*, 11(2), 2018.

[Pratesi *et al.*, 2020] Francesca Pratesi, Lorenzo Gabrielli, Paolo Cintia, Anna Monreale, e Fosca Giannotti. Primule: Privacy risk mitigation for user profiles. *Data & Knowledge Engineering*, 125, 2020.

[Ruggieri *et al.*, 2020] Salvatore Ruggieri, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Dino Pedreschi, e Franco Turini. Opening the black box: a primer for anti-discrimination. In *Annuario di diritto comparato e di studi legislativi*, 2020.

[Setzu *et al.*, 2021] Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, e Fosca Giannotti. Glocalx - from local to global explanations of black box AI models. *Artif. Intell.*, 294:103457, 2021.

[van den Hoven *et al.*, 2021] Jeroen van den Hoven, Giovanni Comandé, Salvatore Ruggieri, Josep Domingo-Ferrer, Francesca Musiani, Fosca Giannotti, Francesca Pratesi, Marc Stauch, e Iryna Lishchuk. Towards a digital ecosystem of trust: Ethical, legal and societal implications. *Opinio Juris in Comparatione*, 2021, 2021.