Data Processing: Reflections on Ethics

Donatella Firmani[1], Letizia Tanca[2], and Riccardo Torlone[1]

[1] Roma Tre University, Rome, Italy
{donatella.firmani, riccardo.torlone}@uniroma3.it
[2] Politecnico di Milano, Milan, Italy
letizia.tanca@polimi.it

Abstract. Ethics-related aspects are becoming prominent in data management, thus the current processes for searching, querying, or analyzing data should be designed is such a way as to take into account the social problems their outcomes could bring about. In this paper we provide reflections on the unavoidable ethical facets entailed by all the steps of the information life-cycle, including source selection, knowledge extraction, data integration and data analysis. Such reflections motivated us to organize the First International Workshop on Processing Information Ethically (PIE).

1 Introduction

Information management naturally involves ethical concerns about how data can be used or misused, posing new challenges to researchers and practitioners across the whole spectrum of information systems. Data is the bridge between stark hardware and people, thus the data produced by any information system cannot convey the appropriate knowledge if humans do not give it semantics. Today, it is widely accepted that a decent system must manage data in a truthful, accurate and secure way, hence the satisfaction of ethical requirements is fundamental in modern applications.

In this paper we discuss the role of the most common ethical requirements of any information system – namely *fairness, transparency, diversity* and *data protection* – and provide reflections and open problems related to how each of them can be considered in the following phases of the *information life-cycle* [17]:

1. *source selection*, i.e., identifying the datasets, or *data sources*, containing the data of interest,
2. *data integration*, namely, extracting and integrating those data in order to produce a unique dataset, and
3. *information extraction*, that is, applying the information extractions tools, from a basic query up to a sophisticated machine learning method, to produce knowledge.

PIE 2019, June 4, 2019, Rome, Italy. Copyright held by the author(s).

Specifically, we discuss how each of these steps may imply ethically relevant choices, advocate the ethics by design in the information life-cycle, and discuss related goals and challenges. We also reason on how a goal might be interdependent with another one.

2 Ethical Requirements

We first briefly review the above-mentioned ethical principles.

- **Fairness** of data is defined as the *lack of bias*, and different notions of bias can yield different fairness measures [16]. Fairness has often been studied for *processes* [12], while recently its importance has been acknowledged also for the data involved in the process itself, due to the (possibly dramatic) consequences of training Artificial Intelligence (AI) systems with biased data, both in the general setting [11] and in specific data management tasks [5].
- **Transparency** is the ability to interpret the information extraction process in order to verify which aspects of the data determine its results. Transparency metrics can use the notions of (i) *data provenance*, by measuring the degree to which the meta-data describe where the original data come from; (ii) *explanation*, by describing how a result has been obtained. Transparency enables the detection of possible biases, thus is somehow "of service" to fairness.
- **Diversity** is the degree to which different kinds of objects are represented in a dataset. Several metrics are proposed in [7]. Ensuring diversity at the beginning of the information extraction process may be useful for enforcing fairness at the end. However, note that sometimes diversity may conflict with fairness [15], for instance when an effort to guarantee diversity leads to loosing sight of the objective merit of the involved people.
- **Data Protection** concerns the ways to secure data, algorithms and models against unauthorized access. Defining measures for privacy can be an elusive goal since, on the one hand, anonymized datasets that are secure in isolation can reveal sensible information when combined [18], and on the other hand, robust techniques such as ϵ-differential privacy [8] can only describe the privacy impact of specific queries. Needless to say, data protection may conflict with transparency.

3 Discussion

We now discuss our viewpoint of the main ethically-relevant aspects for each of the information life-cycle steps.

Phase 1: Source selection. Data typically come from multiple sources, and it is most desirable that each of these complies with the fairness requirement individually. Unfortunately, often sources are biased with respect to some categories. For instance, a source with restaurants in Rome may over-represent restaurants with Italian cuisine. It is thus appropriate to consider ethics throughout multiple sources, so that the bias towards a certain category in a single source can be eliminated by adding others with opposite biases. Another fundamental challenge in the context of source selection is data protection. While fairness can benefit from multiple sources, data shall be protected at the level of the single data source, as adding more information can only lower the protection level, or, at most, leave it as it is. For instance, the case study in [18] about a dataset released by the New York City Taxi showed that with only a small amount of auxiliary knowledge, an attacker could violate privacy of passengers identifying where an individual went, how much they paid, and weekly habits.

Phase 2: Data Integration. Data integration usually involves three main steps: (i) schema matching, i.e. the alignment of the schemata of the data sources, (ii) entity resolution, i.e., identification of the items stored in different data sources that refer to the same entity, and (iii) data fusion, i.e., construction of an integrated database over the data sources, obtained by merging their contents.

For similar reasons as to those discussed for the source selection phase, entity resolution across several data sources owned by different parties can reveal sensitive information about these entities. Examples range from public health surveillance to crime and fraud detection, and national security. We refer the reader to [19] for a survey of existing techniques and challenges of privacy-preserving entity resolution in the context of Big Data. As for the steps of schema matching and data fusion, we observe that groups treated fairly in the sources can become over- or under-represented as a consequence of the integration process, since combining data coming from different sources might lead to the exclusion of some groups. Similar issues arise in connection with diversity.

In all the above steps transparency is critical since, while data provenance and explanation of the intermediate results play key roles in the enforcement of ethical requirements, both can conflict with data protection requirements. The idea of data transparency without privacy violation is put forward in [4], envisioning the application of blockchain technology.

We finally observe that the rise of machine learning and deep learning techniques for the data integration tasks [6] poses new challenges in grasping how integration outputs are produced [20]. We refer the reader to [5,9] for two of the first attempts in explaining deep learning systems for the entity resolution task.

Phase 3: Information Extraction. This step aims at presenting the user with data organized as to satisfy their needs. As an example, among all the possible means for extracting information we focus on two, namely *search* and *aggregation*.

Search is a widely studied task and literature is rich with methods for maximizing user satisfaction [2]. However, in a job candidate selection or for university admissions, as part of the task, we would like to find rankings that also satisfy certain notions of fairness, for instance that different demographic groups be equally represented in the top search results. Interestingly, we believe that the system should allow the possibility *to specify the desired type of fairness*, so that the data scientists can comply with the requirements coming from the customers who, in their turn, will take responsibility for their choices. We refer the reader to [15] for one of the first attempts to define fairness of exposure in ranking.

A notable phenomenon that instead occurs during aggregation is the *Simpson's paradox*, where trends appearing in different groups of data can disappear or even be reversed when these groups are combined.[3] The work of [14] provides a framework for incorporating fairness for specific aggregations based on independence tests, whereas detecting bias in combined data with full-fledged query systems is still an open problem.

4 Further Readings

An early attempt to consider ethics in the broad information systems and data management area can be found in the data quality book [3] by Batini and Scannapieco. In their work, the authors describe a framework for data quality clusters and dimensions, where those in the *trust* cluster are related to some ethics principles, including, *believability*, *reliability*, and *reputation.* Recent developments of these dimensions triggered by the Big Data challenge are discussed in [10].

One of the first attempts to consider the ethical principles as first-class citizens in data management is in the tutorial [16] by Stoyanovich et al., with the primary goal of drawing the attention of the data management community to the emerging subject of responsible data management. The tutorial gives an overview of existing technical work, primarily from the data mining and algorithms communities, and discusses related research directions.

More recently, the work [17] advocates the injection of ethical principles into the whole information extraction process, by properly amalgamating and resolving contrasts between various ethical requirements. The paper provides the vision of a large group of data management researchers towards the description of a comprehensive checklist of ethical desiderata for information processing, to ensure and verify that ethically motivated requirements and related legal norms are fulfilled throughout the data selection and exploration processes.

With an analogous standardization spirit than [17], but in the more specific scenario of building AI systems and supporting robust data analysis, the work in [13] describes the *Dataset Nutrition Label*, that is a diagnostic framework to provide a distilled yet comprehensive overview of dataset "ingredients" for AI model development. Future directions for the project include research and public policy agendas to further advance consideration of the concept.

Finally, Abiteboul et al. [1] bring regulatory frameworks – such as the European Union's General Data Protection Regulation (GDPR), the New York City

[3] One of the best-known examples of such paradox is a study of gender bias among graduate school admissions to University of California, Berkeley. Overall, men were more likely than women to be admitted. However, by examining the individual departments, instead, women were significantly more successful than men. A deeper analysis showed than women tended to apply to competitive departments with low rates of admission, thus yielding the inverse overall trend.

Automated Decisions Systems (ADS) Law, and the Net Neutrality principle – to the attention of the data management community. Governments are starting to acknowledge the importance of building norms and codes for data-driven algorithmic technologies, and such regulatory frameworks are prominent examples. The main take-away of the paper is that in order to comply with regulatory frameworks we shall think in terms of ethics by design, viewing ethics as a systems requirement, rather that incorporating it into systems in retrospection.

5 Concluding Remarks

Ethics in Information Management is a multifaceted concept, including some requirements implicitly related or in conflict. As ethics by design is starting to be recognized as a system requirement, we argue that a key ingredient is to achieve an explicit and holistic vision of ethics as a first-class citizen for data management, as data is at the core of modern information systems. To this end, we discussed the challenges of introducing ethics during the three phases of the information life-cycle, as a necessary step to allow different stakeholders to enact law regulations and equally benefit from modern data processing techniques.

References

1. Serge Abiteboul and Julia Stoyanovich. Transparency, fairness, data protection, neutrality: Data management challenges in the face of new regulation. *To appear in the ACM Journal of Data and Information Quality (JDIQ)*, 2019.
2. Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *2nd ACM International Conference on Web Search and Data Mining (WSDM)*, pages 5–14, 2009.
3. Carlo Batini and Monica Scannapieco. *Data and Information Quality - Dimensions, Principles and Techniques*. Data-Centric Systems and Applications. Springer, 2016.
4. Elisa Bertino, Ashish Kundu, and Zehra Sura. Data transparency with blockchain and AI ethics. *To appear in the ACM Journal of Data and Information Quality (JDIQ)*, 2019.
5. Vincenzo Di Cicco, Donatella Firmani, Nick Koudas, Paolo Merialdo, and Divesh Srivastava. Interpreting deep learning models for entity resolution: an experience report using LIME. In *2nd International Workshop on Exploiting Artificial Intelligence Techniques for Data Management @ SIGMOD*, page 8, 2019.
6. Xin Luna Dong and Theodoros Rekatsinas. Data integration and machine learning: A natural synergy. In *2018 International ACM Conference on Management of Data (SIGMOD)*, pages 1645–1650, 2018.
7. Marina Drosou, HV Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. Diversity in big data: A review. *Big data*, 5(2):73–84, 2017.
8. Cynthia Dwork. Differential privacy. *Encyclopedia of Cryptography and Security*, pages 338–340, 2011.
9. Amr Ebaid, Saravanan Thirumuruganathan, Walid G. Aref, Ahmed K. Elmagarmid, and Mourad Ouzzani. EXPLAINER: entity resolution explanations. In *35th IEEE International Conference on Data Engineering, ICDE*, pages 2000–2003, 2019.

10. Donatella Firmani, Massimo Mecella, Monica Scannapieco, and Carlo Batini. On the meaningfulness of "big data quality". *Data Science and Engineering*, 1(1):6–20, 2016.
11. Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. Ai4people-an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707, 2018.
12. Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *11th ACM Joint Meeting on Foundations of Software Engineering*, pages 498–510, 2017.
13. Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*, 2018.
14. Babak Salimi, Johannes Gehrke, and Dan Suciu. Bias in OLAP queries: Detection, explanation, and removal. In *2018 International ACM Conference on Management of Data (SIGMOD)*, pages 1021–1035, 2018.
15. Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *24th ACM International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 2219–2228, 2018.
16. Julia Stoyanovich, Serge Abiteboul, and Gerome Miklau. Data, responsibly: Fairness, neutrality and transparency in data analysis. In *19th International Conference on Extending Database Technology (EDBT)*, 2016.
17. Letizia Tanca, Paolo Atzeni, Davide Azzalini, Ilaria Bartolini, Luca Cabibbo, Luca Calderoni, Paolo Ciaccia, Valter Crescenzi, Juan Carlos De Martin, Selina Fenoglietto, Donatella Firmani, Sergio Greco, Francesco Isgrò, Dario Maio, Davide Martinenghi, Maristella Matera, Paolo Merialdo, Cristian Molinaro, Marco Patella, Roberto Prevete, Elisa Quintarelli, Antonio Santangelo, Andrea Tagarelli, Guglielmo Tamburrini, and Riccardo Torlone. Ethics-aware data governance (vision paper). In *26th Italian Symposium on Advanced Database Systems (SEBD)*, page 49, 2018.
18. Anthony Tockar. Riding with the stars: Passenger privacy in the nyc taxicab dataset. *Neustar Research, September*, 15, 2014.
19. Dinusha Vatsalan, Ziad Sehili, Peter Christen, and Erhard Rahm. Privacy-preserving record linkage for big data: Current approaches and research challenges. In *Handbook of Big Data Technologies*, pages 851–895. Springer, 2017.
20. Xiaolan Wang, Laura Haas, and Alexandra Meliou. Explaining data integration. *Data Engineering Bulletin*, 41(2), 2018.