Algorithmic Accountability and the Role of Provenance

Jacqui Ayling, Yushi Zou, and Adriane Chapman

University of Southampton
Southampton SO17 1BJ
{j.a.ayling, yz4g16, adriane.chapman}@soton.ac.uk

Abstract. Algorithms are being used in increasingly 'mission ciritcal' applications which make decisions that affect people's lives. Complex algorithms result in a 'black box' from which it is impossible to extract an intelligible explanation of the reasons for an output. This leaves users and regulators with little basis for redress for harmful or unjust outcomes. This paper explores the possibilities of leveraging the techniques of provenance to capture hitherto undocumented aspects of data use in algorithms, contributing to greater accountability for end users.

Keywords: Algorithmic Accountability, Provenance

1 Introduction

As the volume and scope of data collection increases, so does the use of algorithms to analyse, interpret, predict and decide. For the most part these processes are benign and routine, but as the complexity and power of the predictive models of algorithms grows, (especially those that learn from and respond to their inputs), so do the possible dangers. Increasingly 'mission critical' use cases employ algorithms to make decisions, e.g. self-driving cars [3], predictive policing, social welfare [4], medical diagnosis [15], recruitment [10], political campaigning [6] and 'smart city' systems [18]. Any domain where it is relatively easy to capture data is ripe for exploitation, with decisions previously made by human agents now being delegated to algorithms, either completely, or as a source of information to assist in a decision. There is growing concern in the public realm around questions like: 'What is happening to my personal data?' 'How do I argue with decisions made by machines about me?', 'What happens if the machine goes wrong?', 'Who do we blame if it does go wrong?'

Complex machine learning algorithms result in a 'black box' of such opacity that there is no intelligible way to unpack the processes that lead to a particular output from a specified input. The call for Algorithmic Accountability (AA) from both the public and governments therefore becomes a very difficult 'ask'. If a human agent makes a decision then they can be questioned as to the reasons for their decision, and ultimately held accountable if they have failed act as would reasonably be expected under the circumstances [28]. An adequate explanation of a complex algorithms decision-making processes that would make sense to anyone, from a computer scientist to a member of the public, government or legislature, raises serious issues of interpretability [17]. The inherent uncertainty and inscrutability of the processing results in an inability to explain why a particular decision was made. [22]. The forthcoming General Data Protection Regulation (GDPR) includes the right to an explanation of automated decision making and profiling [25, Article 4] but it is as yet unclear exactly how this might be meaningfully achieved.

This paper explores the possibilities of leveraging the techniques of provenance in capturing 'the origins and the history of data in its life cycle' [9], which could be used to address hitherto undocumented aspects of data use in algorithms, contributing to greater accountability for end users. In this work we:

1. Provide definitions of the major concepts within algorithmic accountability that could be tackled with provenance techniques in Section 2.
2. Identify use cases of interest to for Algorithmic Accountability that could be helped by provenance in Section 3.
3. Discuss the problems, gaps and challenges of using provenance to answer Algorithmic Accountability questions in Section 4.

2 Algorithmic Accountability Definitions

"Accountability implies an obligation to report and justify algorithmic decision-making, and to mitigate any negative social impacts or potential harms" [13]. Various research and governing bodies have discussed concepts that could contribute to Algorithmic Accountability and can be broadly summarised as: transparency, explainability, auditability, fairness, responsibility and accuracy [11, 12, 1]. In this section, we provide an overview of these concepts.

Transparency. If the processes can be inspected by those who wish to regulate, audit, monitor and have redress to algorithmic decision-making, then accountability has been achieved. Several barriers to this approach have been identified: 1. Privacy - exposing data sets will contravene the rights of data subjects; 2. Perverse effects - for example transparency enabling gaming the system, or producing stigmatization; 3. Protecting trade secrets and competitive advantage; 4. Opacity of complex systems - the 'black box' problem where no sensible human explanation can be provided for the processes of an algorithmic system.

Explainability. Even if a process is transparent, it may not be explainable. Algorithms and the data that they produce must be understandable to their users and stakeholders. While many attempts have been made to provide technical explanations to an algorithms' behaviour, or a data artifacts' creation, these do not translate understandable to the *user*, not the system IT, developer or technical administrator. There is no form of sensible redress or ability to regulate systems if descriptions are only comprehensible to technical experts [13].

Auditability. The aim of auditability is to allow both internal and external parties understand and review the behaviour of an algorithm through appropriate access to the system (e.g. provision of APIs), documentation of development processes, history of system changes and updates, training data, input and output data [14]. Auditability supports the delivery of transparency, but faces the same barriers identified for transparency, and explainability [1].

Fairness. Unjust and discriminatory outcomes from algorithmic decisions have been identified across many different domains. Algorithms should not reinforce existing social inequities (particularly based on protected characteristics like race, sex, gender identity, nationality and religion, or proxies for those like location, income, educational attainment), and protect individuals and groups from discriminatory harms [28]. Calculating the error rates and types for different sub-populations, and assessing the potential for differential and possibly unjust outcomes, requires careful design, testing and on-going monitoring of the the behaviour of algorithms 'in the wild'[11].

Responsibility and De-Responsibilization. The human agencies responsible for the outcomes of algorithmic decisions should be clearly defined, '[t]he algorithm did it" is not an acceptable excuse if algorithmic systems make mistakes or have undesired consequences...'[14]. Mittelstadt et. al [22] identify the danger of 'de-responsibilisation'in complex data assemblages where human actors cannot interrogate the decision-making process (opaqueness) and have little control over, or information to judge the reliability of results, which leads to a mentality of 'the computer said so'.

Accuracy. Neither the outcome of algorithmic systems, nor the data upon which the decisions are based are error-free. Accountability for the errors that the system produces, and the harms it may produce for users, should be recorded and made comprehensible to stakeholders and users. Sources of errors should be clearly identified, and mitigation proceedures put in place [1, 28, 14].

3 Algorithmic Accountability Use Cases

Examples exist of algorithms making vital decisions in many domains, such as self-driving cars [3], predictive policing, social welfare [4], medical diagnosis [15], recruitment [10], political campaigning [6] and 'smart city' systems [18]. While these vary with respect to domain, specific machine learning algorithm used, and type of input data, they all follow a set of steps that include: 1. data acquisition; 2. data cleaning; 3. data translation; 4. training and testing; 5. use of the model generated to create a decision. Figure 1 is an example for simple prediction system with machine learning algorithm embedded that showcases these categories. The machine learning model cannot be built without first acquiring training and testing data which is then put through a data processing pipeline that cleans and translates the data for easier processing by the ML algorithm. Before the machine learning algorithm is ever invoked, information is lost during the data cleaning and translation processes. Across the entire series of steps depicted in Figure 1, different processes are easier to introspect into, from a white-box, relational world, to a fully-black-box machine learning algorithm. The variation between white-, grey- and black-boxes creates difficulty with transparency and explainability.

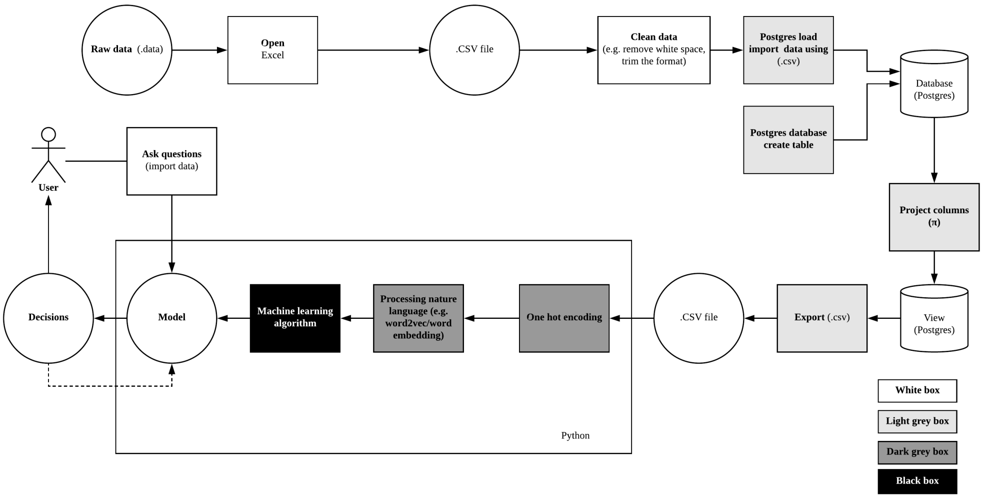
Raw data (.data)
Open
Excel
.CSV file
Clean data
(e.g. remove white space, trim the format)
Postgres load import data using (.csv)
Database (Postgres)
Postgres database create table
User
Ask questions (import data)
Project columns (π)
Decisions
Model
Machine learning algorithm
Processing nature language (e.g. word2vec/word embedding)
One hot encoding
.CSV file
Export (.csv)
View (Postgres)
Python
White box
Light grey box
Dark grey box
Black box

Fig. 1. An example chain of processes and data that ultimately create a decision making algorithm. Processes are categorized into white, grey and black based on the ability to introspect into exactly what occurred within each.

4 Provenance and Algorithmic Accountability Challenges

The concepts identified above that underpin Algorithmic Accountability share similarities to the goals of the provenance community, who use similar concepts as motivation for the overhead of capturing and storing provenance. For example transparency [21, 24], auditability [16], and explainability. [7, 9, 20] However, three large challenges still exist in effectively supporting Algorithmic Accountability needs with provenance. These include:

1. an "impedence mismatch" between what provenance can provide with respect to transparency, auditability and explainability,
2. the current lack of research on using provenance to help with fairness, responsibility, accuracy, and other AA concepts, and
3. the inability to capture appropriate provenance within the tools and algorithms that are of interest.

Impedence Mismatch As [5] notes, aside from intentional secrecy and/or deception, and a wider lack of skills to understand algorithms, calls for greater transparency fail to comprehend "an opacity that stems from the mismatch between mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation.” For instance, Definition 1 describes the traditional way transparency has been used in the provenance community.

Definition 1: ***Transparency*** Given a process, or set of processes, enough information is available in the provenance record to determine the exact set of data and steps that were used to create the output.

Unfortunately, this definition, and its corroloaries for auditability and explainability, lacks any representation of ”human-scale reasoning” a.k.a. a lay-user’s understanding. Instead, any definition that bridges the gap between provenance and the needs for Algorithmic Accountability must contain a representation of the user and their needs, such as Definition 2.

Definition 2: ***Transparency for AA*** Given a process, or set of processes, enough information is available in the provenance record to determine the exact set of data and steps that were used to create the output, *using the semantics and concepts that a given user, U, can understand.*

While there is a body of work looking at ways to assist in user understanding of large provenance graphs [7, 23, 26], additional research is required on how that information is given to the users in a useful and usable manner. [27] is an example of a step forward in this area, but more remains to be done. Moreover, while some work has been done on explainability [2, 8, 19], the answers forthcoming from these systems are not oriented to the correct end user.

Supporting Additional Algorithmic Accountability Concepts The potential for social harm from machine decisions run wild demand that sensible measures be hypothesised and tested to ensure control and oversight. Therefore investigation into techniques that could usefully contribute to regulation of algorithms is required. While it is likely that provenance will not be a silver bullet and cannot sensibly be expected to solve all the issues presented in a complex socio-technical system, further exploration in how to use provenance to support the goal of increasing the levels of accountability, and mitigating harms in algorithmic decision-making systems is required.

Provenance capture and requirements We believe that despite the ‘black box’problem, providing more details on the data sets chosen to make the prediction and the processes applied to those data set(s) is required. In particular, to highlight the differences between the data points used for training and those used for profiling for a particular subject. Brauneis and Goodman [4], for example, describe five types of reasons for excluding data: quality concerns, susceptibility to manipulation, time and place limitations, lack of relevance, and policy considerations other than lack of relevance. An analysis of how these exclusions impact the accountability of the end algorithm needs to be researched.

5 Conclusions

We have discussed the key concepts that will enable robust governance of algorithms (AA) - Transparency, Explainability, Auditability, Fairness, Responsibility and Accuracy. We suggest that provenance capture techniques might contribute to these goals, particularly in throwing light on the data processing undertaken in the training and operation of algorithms. While provenance cannot solve all of the issues present in a complex socio-techincal assemblage, further research is needed to explore it can contribute to improved accountability. There needs to be particular focus on how provenance techniques could support Auditability, whilst also meeting the demands of Explainability, i.e. documenting data use in algorithms described in terms that are comprehensible to a range of users.

References

1. US ACM and EU ACM. USACM - EUACM Statement on Algorithmic Accountability, May 2017.
2. Nicole Bidoit, Melanie Herschel, and Katerina Tzompanaki. Query-Based Why-Not Provenance with NedExplain. In *Extending Database Technology (EDBT)*, Athens, Greece, 2014.
3. Mariusz Bojarski, Philip Yeres, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Lawrence Jackel, and Urs Muller. Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car. *arXiv:1704.07911 [cs]*, April 2017. arXiv: 1704.07911.
4. Robert Brauneis and Ellen P. Goodman. Algorithmic Transparency for the Smart City. SSRN Scholarly Paper ID 3012499, Social Science Research Network, Rochester, NY, August 2017.
5. Jenna Burrell. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512, January 2016.
6. Carole Cadwalladr. I made Steve Bannons psychological warfare tool: meet the data war whistleblower. *The Guardian*, March 2018.
7. Adriane Chapman, Barbara T Blaustein, Len Seligman, and M David Allen. Plus: A provenance manager for integrated information. In *Information Reuse and Integration (IRI), 2011 IEEE International Conference on*, pages 269–275, 2011.
8. Adriane Chapman and H. V. Jagadish. Why not? In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD ’09, pages 523–534, 2009.
9. James Cheney, Laura Chiticariu, and Wang-Chiew Tan. Provenance in databases: Why, how, and where. *Found. Trends databases*, 1(4):379–474, April 2009.
10. Bo Cowgill and Catherine Tucker. Algorithmic Bias: A Counterfactual Perspective. page 3, 2017.
11. Nicholas Diakopoulos. We need to know the algorithms the government uses to make important decisions about us.
12. Nicholas Diakopoulos. Algorithmic Accountability Reporting: On the Investigation of Black Boxes. 2014.
13. Nicholas Diakopoulos. Algorithmic Accountability. *Digital Journalism*, 3(3):398–415, May 2015.

14. Nicholas Diakopoulos, Sorelle A. Friedler, M Arenas, Solon Barocas, M Hay, B Howe, H.V. Jagadish, K Unsworth, A Sahuguet, S. Venkatasubramanian, C Wilson, C Yu, and B. Zevenbergen. Principles for Accountable Algorithms and a Social Impact Statement for Algorithms :: FAT ML.
15. Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, February 2017.
16. Ashish Gehani and Dawood Tariq. Spade: Support for provenance auditing in distributed environments. In *Proceedings of the 13th International Middleware Conference*, Middleware '12, pages 101–120, New York, NY, USA, 2012. Springer-Verlag New York, Inc.
17. Mireille Hildebrandt. Who Needs Stories if You Can Get the Data? ISPs in the Era of Big Number Crunching. *Philosophy & Technology*, 24(4):371–390, December 2011.
18. Rob Kitchin. Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1):14–29, January 2017.
19. Seokki Lee, Sven Köhler, Bertram Ludäscher, editor="Mattoso Marta Glavic, Boris", and Boris Glavic. Implementing unified why- and why-not provenance through games. In *Provenance and Annotation of Data and Processes*, pages 209–213, 2016.
20. B.S. Lerner and E.R. Boose. RDataTracker and DDG explorer. In *Provenance and Annotation of Data and Processes. IPAW 2014. Lecture Notes in Computer Science.*, 2015.
21. Timothy McPhillips, Tianhong Song, Tyler Kolisnik, Steve Aulenbach, Khalid Belhajjame, R. Kyle Bocinsky, Yang Cao, James Cheney, Fernando Chirigati, Saumen Dey, Juliana Freire, Christopher Jones, James Hanken, Keith W. Kintigh, Timothy A. Kohler, David Koop, James A. Macklin, Paolo Missier, Mark Schildhauer, Christopher Schwalm, Yaxing Wei, Mark Bieda, and Bertram Ludäscher. Yesworkflow: A user-oriented, language-independent tool for recovering workflow information from scripts. *International Journal of Digital Curation*, 10(1):298–313, 2015.
22. Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679, December 2016.
23. Luc Moreau. Aggregation by provenance types: A technique for summarising provenance graphs. In *Graphs as Models*, 2015.
24. Leonardo Murta, Vanessa Braganholo, Fernando Chirigati, David Koop, and Juliana Freire. noworkflow: Capturing and analyzing provenance of scripts. In Bertram Ludäscher and Beth Plale, editors, *Provenance and Annotation of Data and Processes*, pages 71–83, Cham, 2015. Springer International Publishing.
25. European Council and Parliament. REGULATION (EU) 2016/679 General Data Protection Regulation, 2016.
26. Cohen-Boulakia Sarah, Biton Olivier, Cohen Shirley, and Davidson Susan. Addressing the provenance challenge using zoom. *Concurrency and Computation: Practice and Experience*, 20(5):497–506, 2007.
27. Andreas Schreiber and Regina Struminski. Visualizing provenance using comics. In *TaPP*, 2017.
28. Andrew Tutt. An FDA for Algorithms. SSRN Scholarly Paper ID 2747994, Social Science Research Network, Rochester, NY, March 2016.