On Evidence Capture for Accountable AI Systems [⋆]

Wei Pang[1,2], Milan Markovic[2], Iman Naja[2], Chiu Pang Fung[1,3], and Peter Edwards[2]

[1] School of Mathematical and Computer Sciences, Heriot-Watt University
Edinburgh, EH14 4AS, UK
w.pang@hw.ac.uk

[2] School of Natural and Computing Sciences, University of Aberdeen
Aberdeen, AB24 3UE, UK
{milan.markovic, p.edwards}@abdn.ac.uk

[3] School of Computing, University of Leeds, Leeds, LS2 9JT, UK
C.P.Fung@leeds.ac.uk

Abstract. This research explores evidence capture for accountable AI systems. First, different scopes of AI accountability are set out by extending existing classification. Based on these scopes, two important and fundamental questions in evidence capture are answered: what types of evidence need to be captured and how we can capture them to facilitate better AI accountability. We hope that this research can provide guidance on building better accountable AI systems with effective evidence capture and initiate further research along this line.

Keywords: Accountability · Artificial Intelligence · Evidence Capture

1 Introduction

Accountability of AI systems has been increasingly studied in recent years, and it has attracted much attention from not only academia [14] and industry [2, 4], but also government [18] and public sectors [11].

Realising accountable AI Systems entails knowing who the people were behind the key decisions made throughout the AI system's life cycle, e.g., how the system was designed and built, how it is being used and maintained, and how the laws, regulations, and standards were followed [10].

A crucial step to achieve this is to capture evidence effectively. To start with, two questions need to be answered: what types of evidence need to be captured and how they can be captured. Answering these two fundamental questions will help implement functional evidence capture components for AI systems,

This research is supported by the RAInS project (https://rainsproject.org/) funded by EPSRC (EP/R033846/1). We thank all other members of the project for their inspiration and suggestions for this research.

thus making AI systems accountable. It will also provide guidance on how we can perform accountability-related investigations (e.g., incident investigation for automated vehicles and bias investigation for AI-assisted recruitment) through effective evidence gathering.

In this research, we will extensively discuss the above two questions. We do not intend to provide specific solutions or frameworks for evidence capture; instead, we aim to provide guidelines and suggestions, and we hope this could inspire further research on this topic.

The rest of the paper is organised as follows: first, different scopes of AI accountability are set out in Section 2. Then based on these scopes, in Section 3 a series of "what" questions are answered. This is followed by Section 4, in which the "how" question is discussed. Finally, Section 5 concludes the paper.

2 The Three Scopes of AI Accountability

AI accountability may have different scopes and meanings in various scenarios. Following the brief discussion in [7], we further extend the following three scopes of AI accountability (which are called the three "senses" of AI accountability in [7]) by providing more details about each scope and expanding the third scope (see Section 2.3). This will allow us to discuss the two questions of evidence capture (what and how) in the following sections.

2.1 Technology-oriented Accountability

In this scope, accountability is considered as a feature or component of an AI system *per se*. An AI system can offer related functions to make itself accountable. These functions include explainability, attributability, auditability, and provenance. Similar to accountability, each of these functions may have different scopes and meanings in various scenarios. Explainability entails enabling the system to justify its outputs (e.g., decisions and predictions). This can be automated by XAI (eXplainable AI) tools, whether model agnostic [12] or model-specific [1]. In the technical context, attributability involves identifying the roles that technical components have, e.g., if the AI System consists of more than one model, then it is important to know which model was responsible for an erroneous result. Auditability entails allowing the system to be inspected and assessed. Provenance entails documenting how the AI system and its components came to be, e.g., the information about where the training data came from, how a model was implemented, and how performance was evaluated.

2.2 Human-oriented Accountability

Within this scope, accountability aims to hold the persons or organisations accountable. This is because the AI systems are made by and for humans (we argue that for the AI systems automatically produced by AutoAI/AutoML [5], humans are the creators of these AutoAI/AutoML systems). This scope of accountability focuses on the persons or organisations who are behind the AI systems, including the AI designers, developers, service suppliers, and users. The proposed Algorithm Accountability Act of 2019 [18] is concerned with the accountability in this scope.

2.3 Systems-oriented Accountability

In the broadest scope, an AI system is viewed as a complex system, for example, a socio-technical system [15] or a tech-legal system [16]. Accountability in this scope involves how one should build an accountable AI system considering not only the complexity from social, technical, ethical, and legal perspectives, but also the complicated interactions of system components across these perspectives. The goal is to build an AI system that is not only technically robust, but also trustworthy and complies with legal and ethical requirements.

It is noted that, further to the classification in [7, 10], we apply a complex system view to this scope, and we consider that an AI system is composed of core AI components and their supporting facilities (e.g., hardware and software), and such an system is operated in, and interacts with its environment.

3 What to Capture

To be accountable for everything means to be accountable for nothing. Correspondingly, capturing everything is neither feasible nor necessary. To decide what we will actually capture (the action), we need to answer the following three questions: what is the scope of capture, what is the capability of capture, and what is the obligation to capture.

First, the scope of capture is determined by the scope of accountability in consideration (as set out in Section 2); we will discuss this in Section 3.1. Second, the capability of capture is subject to both AI system limitations and external constraints; we will address this in Section 3.2. Third, the obligation to capture is often determined by the requirements of specific domains, regulations, laws, and standards; we will cover this in Section 3.3. Lastly, what we will actually capture is the ultimate question, which is affected by the answers of the first three questions. We will discuss this final question in Section 3.4.

In what follows (Sections 3.1 $\sim$ 3.4), we will not produce an exhaustive list of the types of potential evidence in each subsection (as such an exhaustive list is impossible to generate), but rather, we provide the most essential and representative types of evidence, some of which are accompanied by concrete examples.

3.1 The Scope of Capture

Considering the three distinct scopes of AI accountability set out in Section 2, different sets of evidence for capture can be accordingly considered for each scope. We will now discuss them in detail.

Technical Aspect The first scope of AI accountability is concerned with the technical aspect. First, it is essential to record information about the training and evaluation data (e.g., sources, pre-processing processes, and quality analysis) and about the models, which includes the training paradigm and evaluation procedures. Furthermore, explanations of AI predictions and inference processes, fairness, uncertainty, robustness analysis (and even formal verification) for the AI system often need to be recorded for auditing and potential investigations. In many cases, the above information has not been generated or it is not feasible to generate such information beforehand; therefore, whenever possible, the approaches to generating such information should be investigated, initially configured, and documented for post-hoc accountability analysis. For instance, appropriate XAI and fairness analysis tools for the AI system may be prepared and the instructions for using these tools are recorded.

Social and Human Aspect The second scope of AI accountability focuses on human and social activities. Human activities related to the AI system need to be captured in order to hold them accountable. This includes human decision-making processes and human-human interactions, either directly or through AI system components, during the life cycle of an AI system. For example, the following information may be captured as evidence: the stakeholders' meetings and discussions on the AI system to be developed, AI designers' decision making processes on using particular AI models, the interactions between AI designers and developers during the implementation stage of the AI system, and how users operate an AI system deployed in the wild.

Complex System Aspect As for the third and broadest scope of AI accountability, we need to capture not only the information regarding the first two scopes, but also the interactions and information flows of different elements of the complex AI system, including the interactions between people, the AI components, the lower-level software and hardware supporting infrastructure, and the environment which the AI system is operated in and interacts with.

3.2 The Capability of Capture

As mentioned at the beginning of this section, what can be captured is subject to the AI system limitations and other external constraints. As in Section 3.1, we will again consider the different scopes of accountability set out in Section 2 to discuss this in detail.

Technical Aspect The capture ability is determined by the functionalities and limitations of the AI system. The limitations of the AI models being used may affect such ability. For instance, explaining the inference and reasoning processes of black-box models is generally more challenging compared to white-box models. The robustness of some AI models, e.g. some sophisticated deep neural networks, may be hard to analyse against adversarial attacks. For many cutting-edge AI models, their formal verification may be very challenging or even not possible [6].

Social and Human Aspect If the documentation on some decisions made during the AI system's life cycle is not done well or missing, we may not be able to capture related human activities. Considering a legacy AI system, the documentation of which on the design and development stages is missing, we will not be able to capture the activities of the designers and developers as well as their interactions. Therefore, it will be impossible to hold them accountable.

Complex System Aspect Hardware, software, environmental, ethical, and legal factors can all affect the ability to capture. For example, considering the sensors used by an automated vehicle, their limited ability means we can only capture the data up to a certain resolution. Another example that we may not be able to capture some human activities due to privacy and security considerations.

3.3 Obligation to Capture

For a specific AI application, we must consider related laws, regulations, or standards to capture the required types of information or capture them as much as possible. One example is one of the UK national standards for automated vehicles, the BSI standard PAS 1882 [17], which suggests that high frequency/resolution data should be captured 30 seconds before and after an incident involving an automated vehicle, as well as during the incident. Another example is the well-known (and much debated) "right to explanation" of automated decision making in EU's General Data Protection Regulation (GDPR) [13], which demands explanations for decisions made by algorithms.

3.4 Action: What We Will Actually Capture

Having covered the scope, capability, and obligation of evidence capture, we can now discuss what we will actually capture.

It is obvious that deciding what evidence we actually capture should consider the above three factors simultaneously. For a particular application, from a pragmatic perspective, we may start from the obligations, and then examine/improve the capture capability within the scope of capture. By doing this we will get a narrower set of evidence to be captured.

For the above refined set, we propose the following three principles to further refine it: first, evidence capture should not significantly affect the system performance (e.g., accuracy, efficiency, and reliability) or take too much resource (e.g., computational time, storage, and human labour), and we call this *the performance principle*. Second, evidence capture should be less invasive to the AI system and its environment (e.g., requiring no significant change to the AI system or environment), and we call this *the friendly principle*. Third, considering the above two principles, capturing more is better than capturing less, and we call this *the redundancy principle*.

Finally, evidence capturing needs to consider the nature of the application domain and the requirements of the particular accountability investigation. Capturing potential evidence for an automated vehicle is more likely to include hardware and environmental data, such as the vehicle's engine information, road and weather conditions; but capturing potential evidence for an AI recruitment system may focus more on the technical aspect, such as bias analysis and decision/prediction explanation.

4 How to Capture

In this section, we discuss the methods of evidence capture. First, the same three principles in Section 3.4 need to be followed when designing capture methods. Second, evidence capture should be carried out throughout the AI life cycle, including requirement analysis, design, implementation, deployment, operation, and maintenance. Related components and workflows which enable evidence capture should be carefully designed and implemented for each stage of the AI system life cycle. Third, based on the degree of automation, capturing methods can be classified into three categories: automatic, semi-automatic, and manual. We discuss them in detail below.

Automatic capture does not involve human intervention. Google's TFX framework [9] offers functionalities to automatically record machine learning (ML) model training and evaluation information. The sensors of an automated vehicle can automatically collect system and environmental data. Automatic capture can be further divided into two types: passive capture (capture just in case) and active capture (capture initiated by specific events).

Semi-automatic capture requires some degree of human input; for instance, the Model Card Toolkit (MCT) [3], an open-source tool developed for generating the Model Card [8], requires AI developers to manually input some model information, such as the overview, owner, and limitations of the ML model, but MCT can also rely on TFX components to automatically capture information on training data and model performance, and it can automatically generate the final Model Card in HTML format for better inspection. Another example is that knowledge graph has been used to support evidence capture by both human and automatic means [10].

Finally, manual capture is the last resort if the first two approaches are not feasible; for example, gathering stakeholders' meeting minutes and extracting related information from these documents are likely to be done manually.

5 Conclusion

We have extensively discussed two fundamental questions on evidence capture for accountable AI systems: what to capture and how to capture. We hope the discussion can guide more effective evidence capture and thus contribute to the development of better accountable AI systems.

References

1. Agarwal, R., Frosst, N., Zhang, X., Caruana, R., Hinton, G.E.: Neural additive models: Interpretable machine learning with neural nets. arXiv preprint arXiv:2004.13912 (2020)
2. Arnold, M., Bellamy, R.K.E., Hind, M., et al.: FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development **63**(4/5), 6:1–6:13 (2019)
3. Fang, H., Miao, H.: Introducing the model card toolkit for easier model transparency reporting, https://ai.googleblog.com/2020/07/introducing-model-card-toolkit-for.html
4. Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., Crawford, K.: Datasheets for datasets (March 2018), https://www.microsoft.com/en-us/research/publication/datasheets-for-datasets/
5. He, X., Zhao, K., Chu, X.: AutoML: A survey of the state-of-the-art. Knowledge-Based Systems **212**, 106622 (2021)
6. Leofante1, F., Narodytska, N., Pulina, L., Tacchella1, A.: Automated verification of neural networks: Advances, challenges and perspectives (2018), https://arxiv.org/pdf/1805.09938.pdf
7. Millar, J., Barron, B., Koichi Hori, R.F., Kotsuki, K., Kerr, I.: Theme 3: Accountability in AI: Promoting greater societal trust. In: G7 Multistakeholder Conference on Artificial Intelligence. pp. 1–16. Montreal, Canada (2018)
8. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. Association for Computing Machinery, New York, NY, USA (2019)
9. Modi, A.N., Koo, C.Y., Foo, C.Y., et al: TFX: A tensorflow-based production-scale machine learning platform. In: KDD 2017 (2017)
10. Naja, I., Markovi, M., Edward, P., Cottril, C.: A semantic framework to support AI system accountability and audit. In: ESWC 2021. p. in press. Greece (2021)
11. NHS: A guide to good practice for digital and data-driven health technologies (2021), tinyurl.com/NHSAICode, this URL has been shortened
12. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD, San Francisco, CA, USA, August 13-17, 2016. pp. 1135–1144 (2016)
13. Selbst, A.D., Powles, J.: Meaningful information and the right to explanation. International Data Privacy Law **7**(4), 233–242 (12 2017)
14. Shah, H.: Algorithmic accountability. Philosophical Transactions of the Royal Society A **376**(20170362), 20170362 (2018)
15. Shin, D.D.: Socio-technical design of algorithms: Fairness, accountability, and transparency. In: 30th European Regional ITS Conference. p. 205212 (2019)
16. Singh, J., Cobbe, J., Norval, C.: Decision provenance: Harnessing data flow for accountable systems. IEEE Access **7**, 6562–6574 (2019)
17. The British Standards Institution (BSI): PAS 1882:2021 Data collection and management for automated vehicle trials for the purpose of incident investigation. specification, https://shop.bsigroup.com/ProductDetail/?pid=000000000030408477
18. US House of Representatives: H.R.2231-Algorithmic Accountability Act of 2019, https://www.congress.gov/bill/116th-congress/house-bill/2231