AI, Records, and Accountability

by Norman Mooradian, Ph.D.

This article is part of a collaboration between ARMA and AIEF and is included in Information Management Magazine, ARMA-AIEF Special Edition*, which will be available for download in mid-November. A printed version of the special issue will be available as well, for a nominal fee.*

Introduction

Artificial intelligence (AI) promises wide-ranging benefits for society, but it also poses a host of ethical challenges, such as racial and gender bias, liability for harms caused by AI systems, inequality, economic dislocation, and others. The risks and harms posed by AI will have to be addressed at a societal level and at an organizational level. Records management should have a role in a

addressing some of the risks posed by AI given its mission of creating reliable records and its ethical core value of promoting accountability. For years, records management has been helping organizations address emerging legal and ethical challenges such as information privacy, compliance, and eDiscovery. It has expanded its scope, methods, and capabilities to encompass what is now called information governance. In order to address the AI-based issues that organizations will face, the records profession will (1) need to identify the types of problems it is best positioned to address and (2) develop a strategy of evolving its methods to address developments in AI.

To identify ethical issues that records management has a role in addressing and a contribution to make, we can start by identifying its ethical core and competencies, which are enabling accountability and transparency within organizations through the creation and management of trustworthy records. AI-related issues where accountability and transparency are part of the ethical or legal problem should fall within the scope of records management and benefit from its evolving practices. Two prominent examples of such problem areas are: (1) racial and gender bias in AI algorithms and (2) liability for harms caused by AI systems. Both issue areas involve complicated questions of responsibility that require the capture of a reliable and understandable record. Both therefore are good candidate areas for records management to focus on.

To begin to effectively address AI issues, records management needs to address two questions: (1) how to define an AI record in a given context and (2) how to capture an AI record. An AI record is by definition a record of an AI “act” sufficient to document the act and make it intelligible. To address the first question, records professionals should participate in the Explainable AI (XAI) initiative because its goals overlap with records management goals. To address the second question, records professionals should look for lessons in how records can be captured from other systems that share functional components with AI systems.

Definitions of AI

AI refers to computer systems that are able to perform tasks that are considered to require human intelligence – that is, cognitive tasks. Among the common cognitive tasks are reasoning, predicting, planning, understanding, explaining, speaking, perceiving, and learning. The answer to the question of whether AIs instantiate these tasks intrinsically or simply imitate them to achieve their outcomes is in part dependent on one's conception of the human mind and brain. Early AI systems were based in formal/symbolic logic. They used logical languages (their syntax and semantics) to represent domains and generate inferences based on inputs. Expert systems are an example of the symbolic AI approach. They proved difficult to construct and maintain, however, given the limitation of formalistic methods to represent real-world domains whose causal laws and correlations are typically not fully known and describable.

With the explosion of data (big data), statistical approaches to AI found greater success. In particular, the field of machine learning (ML) has grown rapidly, and ML AIs and their constituent learning algorithms have found broad applications. ML AIs learn from their environment and improve their performance over time. ML algorithms operate over data inputs and learn from them in that they refine and develop their representations of the world (their models) in such a way that they can predict outputs based on new inputs, classify inputs, and infer hidden variables. ML algorithms require sufficient data inputs and some form of training. Three types of training approaches are supervised learning, where training data sets include inputs and their correct outputs; unsupervised learning, where training data relies purely on inputs; and reinforcement, in which incorrect outputs are corrected through intervention (Theobald, pp. 18-24). The power of ML and the opacity that results from its adaptation and evolution in relation to the vast quantities of data over which it operates combine to raise or magnify ethical issues in a way that other computer technologies, such as symbolic AIs, did not. XAI, which will be discussed in this article, is an attempt to mitigate the opacity of ML AIs.

Bias

A central ethical issue for AI is bias in ML algorithms. ML algorithms are used widely in services that interface with consumers and citizens. A distinctive feature of ML is its use of statistical methods to analyze big data. It tends to include many more data points on an individual than would be collected using traditional decision methods and/or it uses data points from a broader population. While both points have ethical significance, the first point has privacy implications as well. The second also raises questions about fairness and rights, as information about groups and not the individual is used as the basis for automated decisions made about the individual. Further, of even greater concern is that the information about the groups may be biased. Combining ML with biased information means that the machine can learn to be biased, and the bias can be reinforced by its previous outcomes.

A recent Pew Research Center survey entitled “Public Attitudes toward Computer Algorithms” reported that a majority of people have concerns about the fairness and appropriateness of using AI algorithms to make important decisions about individuals. The report noted that approximately “. . . six-in-ten Americans (58%) feel that computer programs will always reflect the biases of the people who designed them . . .” (Pew, p. 8).

The concerns revealed by Pew are supported by numerous studies. For example, a RAND report cites a study on the use of software used to predict recidivism in parole cases. The algorithms assessed black convicts with a higher risk than nonblack convicts, “. . . even when the nonblack convicts had more severe offenses” (RAND, p. 13). The same report describes how predictive policing software programs over-predict crime rates for certain subpopulations and how the results of skewed predictions become data for the ML algorithms in a vicious feedback loop (p. 15). The skewed predictions raise the issue of biased data (the “data diet”) that ML algorithms process recursively. Unlike rules-based algorithms, ML algorithms cannot be evaluated

and tested at the formal level alone. Rather, the data they process changes their operational principles. This requires assessment based on outcomes and an analysis of the dynamic between the algorithms and their inputs.

Liability

The issue of bias demarcates a broad area of situations in which persons can be harmed through unjust decisions that deny them fundamental goods. These harms are, however, a subset of many other types of harms that can be caused by AI systems. Physical injury, financial loss, and misdiagnosis are only a few broad categories of harm that may result from the implementation of AIs. The well-known case of a fatality caused by a self-driving Tesla is an example of serious physical harm brought about by a type of AI. Whether a drone, an autonomous vehicle (AV), a robotic system, or an information/decision system, AIs often operate in risk-laden contexts. While operating in risk contexts is not new for computer technologies, AIs pose new questions about legal liability that derive from three features of AIs: (1) their autonomy in defining means to achieve their objectives, (2) their ability to learn and thereby evolve their original programs, and (3) the opacity of their internal reasoning processes.

Law and ethics will need to evolve to address liability issues for AIs, just as records management will need to evolve to support law and ethics. A primary way of allocating legal liability (for civil offenses) in the United States is tort law (Smith, p. 12). Two relevant concepts from tort law are negligence and strict liability, and both are used to assess responsibility and assign damages for harms caused unintentionally. Negligence is typically applied to harms caused by humans and employs a “reasonable person” standard to judge culpability. Strict liability, by contrast, is based on causation and requires no fault to find damages. Products liability is a theory of liability that includes strict liability and negligence and has typically been applied to harms caused by computer systems (Ibid). Application of strict liability to AI will likely put emphasis on the

inherent risks in the design and use of the AI, while application of the negligence standard will look at how feasible it would be to reduce or eliminate the risks. The difficulty in applying either standard to ML AIs is that they are designed to learn from data and thereby evolve, not just process data. As the case of invidious bias demonstrated (and such cases implicate additional areas of laws, in particular civil rights law), well-functioning algorithms can behave badly if they have a poor “data diet.” The question of who bears legal and moral responsibility for harms caused by ML AIs will therefore be a difficult issue to settle going forward. From a records perspective, the challenge will be to capture a sufficient record that documents the internal representations of the AI and the causes of those representations, where those causes will often be the previous data analyzed. For this challenge, the topic of XAI is important.

Explainable AI

XAI is a research focus that attempts to make the decision making of AI systems more transparent and understandable to those using and affected by AI. The initiative has numerous stakeholders, including technical professional associations, regulators, and governmental and private sector AI user organizations. The Defense Advanced Research Agency (DARPA), which is funding research projects in XAI, defines the objectives of its initiative as the development of “. . . new or modified ML techniques that produce explainable models that, when combined with effective explanation techniques, enable end users to understand, appropriately trust, and effectively manage the emerging generation of Artificial Intelligence (AI) systems” (DARPA, p. 5). The European Union High Level Expert Group on AI has explicability as one of its principles of trustworthy AI (HLEG on AI, p. 10). The Institute of Electrical and Electronics Engineers (IEEE) makes transparency in the “. . . internal reasoning processes” of AI systems a technical requirement for safe and beneficial general intelligent systems (IEEE, p. 77).

A paradox of developments in AI is that the most successful approaches (in particular, ML in all its varieties) are the most opaque. Earlier AI systems, such as expert systems, were rules based and expressed in formal logic systems that were in principle understandable by humans. Models were based in if-then conditionals, decision trees, and ontologies, wherein relations between categories of things were represented. As noted earlier, these systems were limited and difficult to construct. ML algorithms, coupled with big data inputs, have proven to be very powerful but have also been opaque at the operational level. Some ML AIs can even form different models (hypotheses about correlations between things and their features) when fed different data. For this reason they have been called “black box” algorithms, and the inscrutability of their decisions is viewed as a barrier to trustworthy computing.

Techniques being explored in XAI to make ML less opaque include developing ML AIs that generate more explainable features or representations or that use more interpretable models, or to develop algorithms that can infer explainable models from black box algorithms (DARPA, pp. 7-8). The latter technique might be considered to involve meta-algorithms that produce “theories” of other ML algorithms. An important component of the DARPA research initiative is to develop interfaces between the explainability functionality and human users such that users can interrogate AI systems as to the basis of their decisions. This component of the initiative is of direct relevance to records management, as interface development should take into consideration the production of reliable records that can be captured by user organizations.

Strategies for Records Management

Records management has evolved to meet ethical and legal challenges posed by technological developments (e.g., eDiscovery, information privacy). It will similarly evolve to support accountability in the area of AI, though challenges

will remain. To do so, it will need to develop new approaches, concepts, and methods, but experience from other expansions can be drawn upon. The first step is to define an AI record – that is, to define the scope of records needed to support accountability. This will be an ongoing and evolving task, but a clear definition of the documents and data that constitute an adequate record is a prerequisite to any effective records management practices in the area of AI. The second step is to develop practices for capturing the full scope of records. Having established criteria for a sufficient AI record provides a normative standard. Capturing the information identified in the standard will be a challenge and will require interdisciplinary teams. This too will be an ongoing initiative, but one in which records professionals should be key subject matter experts and stakeholders.

In defining a sufficient AI record, its scope and contents needed to be characterized. Its scope will be actions, transactions, and events that are carried out (fully or in part) by AI algorithms. Its scope therefore potentially can be as inclusive as any records program insofar as AI algorithms infuse organizational actions, transactions, and events, many of which are already computer mediated. Further, as AI and IoT (internet of things) expand the range of actions, transactions, and events carried out by organizations, the scope of organizational records will increase and along with it the scope of AI records.

As regards the contents of an AI record, a few target areas should be considered at the outset. First, it can be expected that as the regulatory environment changes, compliance documentation will be required for AI implementations. Just as data systems that capture personal information require privacy impact assessments (PIAs) in certain jurisdictions, ethical impact assessments are likely to emerge as a kind of compliance record. Basic compliance documentation should be captured as a record series and referenced by the AI record. PIAs are a good example of the kind of compliance documentation that may be required by law or best practice in the

future (and are already required in many jurisdictions for AI algorithms that process personal information). PIAs require a description of the technology, its use cases, and risks attendant upon its application, as well as mitigation plans. For AI-enabled actions, transactions, and events, a record of these should reference the controlling compliance documentation in place when they happened.

Second, AIs consist of algorithms and other technical structures, so base system design and testing documentation for any implementation should be part of the record. ML AIs are more than their designs, of course, but the design documentation is a foundation. As with the compliance documentation, the AI record can reference the technical documentation for the algorithms underlying the transactions. The link back to the relevant documentation will need to be as granular as the deployment of the algorithms and will therefore need to be version-specific or iteration-specific.

Third, and most challenging, records of decisions need to be captured. These records will constitute the bulk of the AI record and will be transactional or case file record sets. They will consist of summaries of the algorithms deployed in relation to the decision, the data processed, and the internal representations of the AI during the processing. Capturing a record of the decision process and representing the data used as inputs are technical challenges tied to the goal of XAI. As described earlier, XAI aims to make specific decisions explainable through an interface that allows users to interrogate the decision or have the system present a summary of its reasoning. For ML AIs, the technical challenge is substantial. Records professionals will depend on developments in XAI to be able to create usable records. Nevertheless, they should play an active role in shaping requirements for XAI in relation to records such that usable records are capturable. It may be necessary for the profession to develop a specialization for records

professionals who work on interdisciplinary teams of data scientists and other IT professionals, but, as noted earlier, the records profession has had to evolve with technological change, so further evolution would be in keeping with its recent history. In any event, XAI has the potential to lay the basis for sufficient AI records, but the records profession and records professionals are essential to the development of guidelines for records that can be relied upon in legal and other proceedings where records are scrutinized and tested.

ML AI implementations will be increasingly common in the near future. Instead of waiting for the adoption of such technologies, records professionals should begin by assessing how they currently capture records of decision systems. They should also review whether their organization processes big data and how it captures a record of its uses. In the case of decision systems, enterprise-wide data and content management systems often have rules-based workflows that include decision points. As part of the implementation of such workflows, it should be possible to capture a record of the configuration or programming of the workflow as well as an audit trail of the key decisions and actions executed in any workflow. Records professionals should be participating in the requirements process for decision systems to ensure that reliable, complete, and usable records are created. Doing so will address current needs but also serve as a preparation for future AI deployments.

In the case of big data (e.g., social media, IoT data, or other big data sources), organizations should review how records are captured and managed. Big data can create a deluge of information coming into an organization, and for management purposes this data may need to be purged in short-term intervals if not immediately. Assessing record needs in relation to big data flows is critical to organizational accountability, however. Where usable records are required, a balance may need to be found between raw data and syntheses or summaries of the data that is manageable. Developing feasible means of capturing big data records can be an answer to real and present needs within the organization, and at the same time it will serve as a preparation for

capturing a record of data inputs that will be used by ML algorithms in areas where risk of unfair bias and harms is present. In sum, the benefit of evaluating current records in relation to decision systems and big data is that records professionals can start building capacity in advance of AI implementations and can also address present gaps relative to current decision/information systems while doing so.

Conclusion

This article reviewed ethical and legal risk areas arising from AI where the need for reliable, authentic, and usable records is a necessary condition for addressing those risks. It argued that the ethical core of the records profession – namely, enabling accountability in organizations, and its core competencies of defining and capturing records from diverse content types – makes records an important field in and contributor to the emerging interdisciplinary effort to govern AI technologies. The central risk areas reviewed were bias (e.g., racial and gender) in AI algorithms and liability for harms caused by AIs. The risk areas reviewed are broad but not exhaustive. Other types of ethical and legal risk will arise that will require accountability and, by implication, the ability to capture records. The records profession can play an important role in mitigating risks and harms arising from AIs, but it will need to expand its toolkit to do so. Defining an AI record and developing methods for capturing AI records is a project the profession should take on. Joining cause with XAI initiatives is a good place to start. Identifying gaps in the current state of records programs in relation to automated decision systems and big data is another step that can be taken in tandem. The records profession has been responding to challenges in organizational transparency brought on by technological developments for a number of decades and has evolved and expanded in the process. AI presents a new set of challenges as well as new opportunities that one can reasonably expect will be met and seized upon by records professionals.

References

Abbott, Ryan. 2018. “The Reasonable Computer: Disrupting the Paradigm of Tort Liability.” The George Washington Law Review. 89 no. 1

DARPA, 2016. “Broad Agency Announcement: Explainable Artificial Intelligence (XAI).” DARPA-BAA-16-53 (https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf)

High Level Expert Group on Artificial Intelligence. 2018. “Draft Ethical Guidelines for Trustworthy AI”. European Commission. (https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_draft_ethics_guidelines_18_december.pdf)

IEEE. 2018 Ethically Aligned Design. Version 2. IEEE. (https://ethicsinaction.ieee.org/)

Osoba, Osonde, William Welser IV. 2017. “An Intelligence in our Image: The Risks of Bias and Errors in Artificial Intelligence.” RAND (https://www.rand.org/pubs/research_reports/RR1744.html)

Smith, Aaron. 2018. “Public Attitudes toward Computer Algorithms.” Pew Research Center. (http://www.pewinternet.org/2018/11/16/public-attitudes-toward-computer-algorithms/)

Theobald, Oliver. 2017. Machine Learning For Absolute Beginners: A Plain English Introduction, 2nd Edition. Palo Alto, CA: Scatterplot Press.