Working in contexts for which transparency is important:
A recordkeeping view of Explainable Artificial Intelligence (XAI)

Purpose

This paper introduces the topic of Explainable Artificial Intelligence (XAI) and reports on the outcomes of an interdisciplinary workshop exploring it. It reflects on XAI through the frame and concerns of recordkeeping.

Design/methodology/approach

This paper takes a reflective approach. The origins of XAI are outlined as a way of exploring how it can be viewed and how it is currently taking shape. The workshop and its outcomes are briefly described and reflections on the process of investigating and taking part in conversations about XAI are offered.

Findings

The article reinforces the value of undertaking interdisciplinary and exploratory conversations with others. It offers new perspectives on XAI and suggests ways in which recordkeeping can productively engage with it, as both a disruptive force on its thinking and a set of newly emerging record forms to be created and managed.

Originality/Value

The value of this paper comes from the way in which the introduction it provides will allow recordkeepers to gain a sense of what XAI is and the different ways in which they are both already engaging and can continue to engage with it.

Keywords: Explainable Artificial Intelligence, Accountability, Transparency, Recordkeeping, Interdisciplinarity

Article Classification: Viewpoint

Introduction

Artificial intelligence (AI) is not a new topic in human imagination, although many date its origins as a modern phenomenon to 1956 and a workshop held at Dartmouth College, USA. Since that time, artificial intelligence has gone through many phases of development, coming in and out of fashion through a number of so-called AI winters. This history shows the truth of a statement made in this journal over thirty years ago by Ralph Cornes; "Artificial intelligence (AI) which many saw as the wave of the future will arrive by osmosis. Other branches of IT will steal its clothes. It is already starting to happen" (Cornes, 1989). This osmosis makes AI quite challenging to define, for it acts as a broad descriptor for a number of other techniques and terms from expert systems to algorithms to machine learning. This osmosis also means that AI has crept up on us by stealth. Whether we like it or not, it now permeates through our technology enabled lives and many of the systems we use and rely on within healthcare, security, insurance and commerce.

A recent consideration of AI in the archive provides a helpful categorisation of AI that applies a distinction between rule-based systems, statistical models, and deep learning models. This framework is introduced in terms of the way in which expert (human) knowledge is encoded, but another variable it highlights is the relative opacity of the type of systems that result (Rolan *et al.*, 2019). For example, in comparison to rule-based systems, where "inspection of their internal processing is straightforward" and statistical models that "provide a built-in capacity for reporting confidence in a given outcome", deep learning models "result in 'black-box' systems with workings that are difficult to interpret" (Rolan *et al.*, 2019). The evolution of AI and the increasing opacity of the models it employs has not gone un-noticed, and within the AI community a new area of interest has arisen in recent years around the topic of Explainable Artificial Intelligence. For example, sessions and workshops on Explainable Artificial Intelligence or XAI have been held at: the International Conference of Machine Learning (2016); the Annual Conference on Neural Information Processing Systems (2016); the International Joint Conference on Neural Networks (2017); and the International Joint Conference on Artificial Intelligence (2017). There is also a major DARPA (Defense Advanced Research Projects Agency) funded research project on the topic currently in progress (Turek, 2018).

One definition of XAI is that it is: “the challenge of shedding light on opaque machine learning (ML) models in contexts for which transparency is important” (International Joint Conference on Artificial Intelligence, 2017). This definition starts to establish a connection between XAI and recordkeeping, as recordkeeping also operates in contexts for which transparency is important. Indeed much effort has been devoted within the recordkeeping community to highlighting the vital importance of transparency to the proper functioning of societies. This paper reflects on an attempt by the author to engage with XAI from her perspective as a recordkeeper. It starts with discussion of how XAI is viewed from that perspective and the different forms in which it is possible to engage with it. It highlights a distinction between explainability and interpretability as important in defining XAI. It continues to report on a workshop in which the author sought to bring together individuals from a range of backgrounds to discuss the topic further. Following on from the workshop, the author continued to reflect on the process and her resulting attempts to map out a common ground and re-locate recordkeeping on that new territory conclude the article. Through this process, the common ground forms around concepts of not just transparency, but also accountability, fairness, social justice, and trust and a specific focus emerges on the concept of explainability.

Engaging with (X)AI from a recordkeeping perspective

The increasing use of more opaque AI techniques is generally framed as disruptive for recordkeeping. For example, statements have been made that; “The opacity of AI algorithms directly impacts the kind of recordkeeping that may be performed in relation to transactions driven by such technologies” (Rolan *et al.*, 2019). This opacity impacts on recordkeeping in that it brings into question what the record should look like and what it is that we should capture that would make this opacity more transparent. As The National Archives in the United Kingdom has put it: “The uncertain and unbounded nature of new forms of records, such as those derived from machine learning systems, is causing us to rethink how we preserve evidence of these systems, and what is the ‘public record’ that we are preserving” (The National Archives, 2018). One assertion of this article is that the disruption AI causes for recordkeeping arises in large part because of the way it highlights the recordkeeping field’s own (often implicit) biases in applying its own terms, most notably that of record. That is to say the disruption comes as the field is forced to rethink the record and its role in it. If business is no longer to be transacted only by human beings, but also by AI agents, or some combination of the two, what will evidence of those transactions look like, what will the record be? The way in which XAI might cause recordkeepers to rethink will be returned to later. First though, if those like the author, coming from a background in recordkeeping theory and practice, are to engage usefully with XAI, they need to be clear about what it is with which they are seeking to engage. They need to decide how to view it.

One way in which they can view it is as a highly technical research field, one in which recordkeepers are not actively engaging and indeed face barriers to doing so, e.g. in terms of a lack of the required level of expertise for entry. A position paper presented at the Sixth International Conference on Learning Representations characterises this research field as having a "research and publication culture that emphasizes *wins*, most often demonstrating that a new method beats previous methods on a given task or benchmark" (Sculley *et al.*, 2018). One of the authors of this paper had previously given a presentation to the 2017 Annual Conference on Neural Information Processing Systems in which he compared machine learning to alchemy as a way of suggesting to his peers that there was a lack of incentive for developing empirical rigor and deep theoretical understanding within this culture (Pfeffer, 2018). Put simply, a certain vagueness about why a particular algorithm or model worked better than another one could be overlooked so long as the win could still be demonstrated; explanation was less important than performance against certain arbitrarily drawn benchmarks in very specific tasks.

Understanding this culture is key to understanding a distinction which is increasingly being used to bring shape to the new field of XAI, namely that between explainability and interpretability. Gilpin et al. have attempted to make this distinction in the following way; by associating explainability with "models that are able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions" and interpretability as "loosely defined as the science of comprehending what a model did (or might have done)" (Gilpin *et al.*, 2018). This distinction between interpretability and explainability is also mirrored in the European Panel report mentioned earlier, which seeks to distinguish between understanding "how the system works" and "how it behaves" (Panel for the Future of Science and Technology, 2019).

Interpretability or comprehending what the model did/how the system works brings more focus to XAI as highly technical activity looking to establish rigor and theoretical understanding in the development of new AI techniques. As such XAI is situated firmly within the wider AI research field as a growing sub-field. For example, Gilpin et al. have recently sought to sketch out a taxonomy with which to provide structure to multiple different methods and approaches to interpretablilty, including linear proxy models, automatic rule extraction, salience mapping, attention networks and disentangled representations (Gilpin *et al.*, 2018).

Explainablity, on the other hand, brings more focus to XAI as a new and more interdisciplinary conversation, which is starting to move beyond the confines of the more narrowly technical AI field and to draw on other fields such as philosophy and the social sciences. One recent attempt to map out this new conversation identified nine overlapping communities involved in it including: Early Artificial Intelligence; Intelligent Systems, Agents and User Interfaces; Ambient Intelligence – Sensing and Context-Awareness; Interaction Design and Learnability; Interpretable ML and Classifier Explainers; Algorithmic Fairness, Accountability, Transparency, Policy and Journalism; Causality; Psychological Theories of Explanations; and Education and Cognitive Tutors (Abdul *et al.*, 2018).

It is noticeable that the recordkeeping community do not make this list, but then the same is true of many others. Indeed the more academic interdisciplinary conversation being mapped out above can be seen as a reaction to a growing concern and conversation amongst society at large about the impact of utilising AI techniques in ways that may disproportionately and adversely affect certain individuals and groups. Indeed some recordkeeping practitioners are already facing the task of implementing the General Data Protection Regulation and its provisions in respect of a) what is and is not socially and legally acceptable with regards to automated individual decision-making and profiling, and b) how "meaningful information about the logic" involved in such decisions does not need to mean "over-complex explanations of algorithms" (Information Commissioner's Office). In dealing with this task, those involved in recordkeeping must ensure that they are a part of the wider societal and interdisciplinary conversations that are starting to occur. It is only through such conversation, and in collaboration with others, that answers will be found.

Interdisciplinary conversation

Seeking to create an opportunity for the sort of interdisciplinary conversation discussed above, the author, working with staff members from The National Archives and a researcher from the field of Human Computer Interaction (HCI), organised a workshop at which the topic could be explored alongside additional participants from a range of backgrounds. The workshop was funded by a grant from University College London's Grand Challenges Transformative Technology Small Grants Fund. XAI is an area of interest for those working in HCI as much as it is for those working in recordkeeping and a recent analysis of the XAI literature by HCI researchers led to a research agenda which framed the role HCI could play as follows:

> While researchers in the ML [Machine Learning] and AI communities are working on making their algorithms explainable, their focus is not on usable, practical and effective transparency that works for and benefits people. Given HCI's core interest in technology that empowers people, this is a gap that we as a community can help to address, to ensure that these new and powerful technologies are designed with intelligibility from the ground up (Abdul *et al.*, 2018).

It would seem a safe assumption that the recordkeeping community would be in sympathy with such an agenda and with this focus on "usable, practical and effective transparency that works for and benefits people" the workshop invented a new acronym HeXAI as it wished to explore Human-Centred Explainable Artificial Intelligence (Abdul *et al.*, 2018), (Bunn, 2019).

A call for participation was put out, with the effect that the participants in the workshop were all self-selected. Indeed, in order to ensure as open a discussion as possible, the only limit applied to those who could participate was the size of the room in which it was to take place. Aside from the organisers, there were nine responses received (some from individuals and some from groups of individuals). All those who wished to take part were asked to provide an initial position paper in advance and these are available on the workshop blog (Bunn, 2019a). In the end there were thirteen attendees in total; the four organisers and nine participants. Of the attendees; eleven were research active members of university staff or doctoral or masters students, and two worked at The National Archives. The backgrounds of the attendees varied, e.g. geology, computer science, archaeology, HCI, archives and records management etc., as did their level of knowledge of AI, with the majority of the participants not research active in that field. The workshop was based primarily around more open-ended discussion to see what sort of sense of XAI was built in common by participants during the day and participants were not allowed to introduce themselves. Instead participants were introduced by a partner assigned in advance as a way to initiate the process of active engagement with perspectives and positions other than one's own.

Given the explicit human-centred framing, it was unsurprising that this was where the focus ended up and a sense of XAI emerged that was very much about people and not technology. For example, there was a feeling from those with more knowledge of and history in AI, that AI was now being used as a general label for what they felt was only a tiny part of their field - recent advances in certain deep learning techniques. Returning to the categorisation of AI mentioned earlier, such techniques or models are only one part of over 70 years of research and development, and their particular opacity is not necessarily indicative of all AI methods. There was a need felt for greater public understanding of what AI actually was and the different ways in which it has been realised by the research community who have created it.

Another way in which this human-centred focus emerged was through a sense that we needed to change the metaphor. A powerful metaphor associated with XAI is that of the black box, but this places the technology and the algorithm firmly at the centre of our focus as an unknowable, and perhaps threatening gap, something (a black box) we cannot see into and hence cannot understand, or reason with. An alternative suggested was that of the iceberg. Here we knew that there was more below than the surface than we could immediately see, but we also had agency in choosing to look above or below the water line. In some cases we might not care what was below the surface, but in others it could be very important.

Discussions about a shift in metaphor, reflected another shift that the participants seemed to want to make; away from AI and towards the human need for explanation and insight at specific moments and in specific contexts. Explanation was seen as an iterative and interactive process and also as a contextual human behaviour with a role in cementing social cohesion and trust. It was felt that engaging with XAI meant also engaging with questions such as the following. When do we need to offer an explanation? When do we want to receive one? How detailed do they need to be? What is a good enough explanation in any given circumstance? Why do we need explanations?

Before the workshop, the author would have anticipated that the common ground to emerge in an interdisciplinary conversation about XAI would appear in terms such as transparency and accountability. What was not anticipated was the focus on explanation. Transparency and accountability were there in the background, but it was explanation and explainability that firmly occupied the foreground. Continuing to reflect on the workshop, the author sought to explore both background and foreground in more detail and these explorations are set out in the next two sections.

Mapping out the common (back)ground

The placing of XAI within "contexts for which transparency is important" has already been mentioned and one workshop that has been held since 2014 is titled Fairness, Accountability and Transparency in Machine Learning (www.fatml.org). The resulting FAT becoming a common acronym within both XAI and beyond. A recent report from the European Panel for the Future of Science and Technology seeks to set out a "governance framework for algorithmic accountability and transparency" in the context of fairness, which it describes as "a guiding purpose for transparency and accountability" (Panel for the Future of Science and Technology, 2019). Then again, there are many other sets of principles seeking to guide AI development that also heavily feature accountability and transparency (Independent High-Level Expert Group on Artificial Intelligence, 2019; Organisation for Economic Co-operation and Development, 2019; The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019).

Fairness as an idea is not highlighted as much within recordkeeping literature, but it is explicitly associated (in the Panel report) with another that has been, that of social justice framed as:

> Ideal vision that every human being is of equal and incalculable value, entitled to shared standards of freedom, equality, and respect. […] It specifically draws attention to inequalities of power and how they manifest in institutional arrangements and systemic inequities that further the interests of some groups at the expense of others in the distribution of material goods, social benefits, rights, protections, and opportunities (Duff *et al*., 2013).

Recordkeeping and XAI would seem then to share both an appeal to this ideal vision and a view that transparency and accountability are two of the mechanisms by which it can be achieved. XAI however sits in a context in which this vision is framed much more strongly as fairness and the avoidance of bias. Work has been undertaken that demonstrates that technological systems (built using AI techniques) are not immune to bias and how, for example, search engines reinforce racism, facial recognition systems are better at recognising persons of certain ethnicities and so on (see for example O'Neil, 2016; Noble, 2018). That systems human beings have built have as much potential to be biased, discriminatory and unfair as human beings do should perhaps not come as a surprise. It is however problematic for those who wish to sell AI as instead offering the potential to be different, to be more objective and all-knowing in its application of an artificial (perhaps even divine) intelligence, at once removed from our own flawed and limited humanity.

Returning to transparency and accountability, the two terms do seem to be similarly defined in both recordkeeping and XAI spaces. For example, a set of Principles for Accountable Algorithms developed in 2016 does offer a definition of accountability as "an obligation to report, explain, or justify algorithmic decision-making as well as mitigate any negative social impacts or potential harms" (Diakopoulos *et al.*, 2016). This chimes with the recordkeeping understanding of accountability as defined by the Society of American Archivists' Glossary as, "The ability to answer for, explain or justify actions or decisions for which an individual, organization, or system is responsible" (Pearce-Moses, 2005). Whilst it is interesting to note the different framing of obligation versus ability, there would again seem to be some common meaning and recordkeepers should not find the following definition of accountability too unfamiliar or alien:

> a set of a mechanisms, practices and attributes that sums to a governance structure which involves committing to legal and ethical obligations, policies, procedures and

> mechanism, explaining and demonstrating ethical implementation to internal and external stakeholders and remedying any failure to act properly (Panel for the Future of Science and Technology, 2019).

Moving on to transparency, it would seem that the definition of this term is more taken as read in both XAI and recordkeeping because explicit definitions proved harder to find. One attempt to bring definition to transparency in relation to AI sees it as a tool “to be used responsibly, which means accepting that applying it means being sensitive to the complex contexts in which it is used, and the balance of benefits and harms its use inevitably entails” (Panel for the Future of Science and Technology, 2019). This is arguably, a more complex definition than the following from the InterPARES Glossary that defines it as “(The condition of) timely disclosure of information about an individual’s or organization’s activities and decisions, especially to support accountability to all stakeholders” (Pearce-Moses, 2018). Nonetheless, there would appear to be some common ground in how both XAI and recordkeeping link transparency with accountability; i.e. transparency “is implied by the most basic conception of accountability: if we cannot know what an organisation is doing, we cannot hold it accountable, and cannot regulate it” (Panel for the Future of Science and Technology, 2019).

Rethinking the record and our role in it – the record as explanation

In exploring the common background above, some differences between recordkeeping and XAI perspectives and framing did appear. For example, within definitions, there was a difference between the stance of seeing accountability as an obligation on oneself and that of seeing it as an ability, presumably one you helped to support in others. Then again, another difference was that in the XAI framing on fairness, accountability and transparency, there was no equivalent idea to that of recordkeeping’s authenticity, even though, as was noted, AI is in many ways as forward as recordkeeping in making a claim to objectivity and objective knowledge. Noticing and considering what such differences might mean is always one of the most valuable outcomes to arise from conversing with others about things you do not yet understand. However as has already been stated, the one thing that stood out most for the author in the conversation engendered by the workshop was the emphasis on explanation.

The European Commission High Level Expert Group on Artificial Intelligence has recently released a set of Ethical Guidelines for Trustworthy Artificial Intelligence. Trustworthiness is an idea that also appears in the recordkeeping field with for example, the recent work of the InterPARES Trust project 'exploring issues of trust and trustworthiness of records and data in online environments' (InterPARES Trust, 2018). As with accountability and transparency, common ground with recordkeeping can be built on the terms of trust and trustworthiness. And yet whereas recordkeeping framings tend to decompose trustworthiness into elements of reliability, authenticity and accuracy; the Ethical Guidelines for Trustworthy Artificial Intelligence see the foundations for such a thing in 'respect for human autonomy, prevention of harm, fairness and explicability' (Terminology Cross-domain Task Force, 2008; Independent High-Level Expert Group on Artificial Intelligence, 2019). Explicability is thereby raised to the level of a fundamental ethical principle and presents itself as an alternative focus in efforts to build trustworthiness.

One suggestion for those (such as The National Archives) involved in asking "what is the 'public record' that we are preserving" is to consider how ideas of record as information and evidence can be enhanced by consideration of them as explanation (The National Archives, 2018). For example, there would appear to be some potential in explorations of the meaning of explanation that highlight causal history and causal responsibility. Could not these start to expand our understanding of provenance not as creator, but as creation in the sense of causal history? Then again, what would happen if we started to apply the same framework Miller has recently applied to explanation and start to see the record as a cognitive process (primarily of abductive reasoning), a social process and a product all at the same time (Miller, 2019)? Resolving this sort of disruption, to our thinking about what we are doing, is an exercise in which recordkeepers must take the lead, but the answers will not be simple or arrived at quickly. What then can we do and should we be doing now?

Engaging with XAI – some practical suggestions

In answering this question and deciding what to do immediately, it will not be open to many recordkeeping practitioners to gain advanced understanding of the intricacies of developments in the interpretability of AI techniques. Nevertheless, it is possible, as this article and the workshop have demonstrated, to gain sufficient understanding of the ideas of interpretable and explainable AI in order to have productive conversations with others from both within and without the AI community. Recordkeepers are also well placed to focus on doing what they have always done; gaining an expert sense of what sort of records are being produced and assessing their suitability to act as a permanent public record. This focus could include ethnographic study, of the sort advocated for by Karen Gracy as far back as 2004, within the AI research community (Gracy, 2004). As that community starts to work out a more rigorous base for itself, who better than recordkeepers to offer advice on how that base might be instantiated in a sustainable record form? Then again, moving out into the application environment, could not recordkeepers also try to engage more with existing standards for software engineering, e.g. ISO/IEC/IEEE 12207: 2017 and the forms of record that may already arise from processes such as code reviews, system specification and testing (International Standards Organization, 2017)?

Moving into the more specific XAI and algorithmic accountability space, some new record needs are already being conceptualised. For example, the Association of Computing Machinery's Statement on Algorithmic Accountability talks of how "A description of the way in which the training data was collected should be maintained" and also how:

> Institutions should use rigorous methods to validate their models and document those methods and results. […] Institutions are encouraged to make the results of such tests public (Association of Computing Machinery, 2017).

Then again, taking the first of these ideas further, the European Report on a Governance Framework, starts to flesh out a "Datasheets requirement" in the form of "a semi-structured document that asks questions such as 'Why was the dataset created?,' How was the data collected?" (Panel for the Future of Science and Technology, 2019). It also imagines an "Archive of Systems Decisions" which would document, at a minimum, decisions not to subject certain systems to another imagined record form – an algorithmic impact assessment (Panel for the Future of Science and Technology, 2019). As was noted above algorithmic accountability can be seen in terms of a governance structure. That governance structure will require certain record forms and it is on them that the recordkeepers' focus is perhaps best placed, to prevent any reinventing of the wheel and to ensure that the record forms that are eventually created serve their purpose and serve it well.

Conclusion

Explainable Artificial Intelligence has been introduced in two ways. Firstly it has been introduced as a concern felt within the AI community with the interpretability of the techniques they are developing and with embedding a more rigorous and less win-oriented culture into their research. Secondly as a new and emerging conversation that has arisen as a reaction to growing societal concern about the way in which AI techniques are being deployed in real world applications. This conversation is taking place in both academic and non-academic circles and between individuals from all kinds of backgrounds.

Initiating one such conversation, led to a workshop which took a human-centred view and to questions such as how best to bring to light the humans behind AI and how to change the metaphor to something more giving of human agency. The workshop also placed an emphasis on the idea of explanation as a contextual human behaviour which supported trust and social cohesion.

Following on from the workshop, the author reflected on the conversation had with others from different backgrounds and continued to try to make sense of XAI from a recordkeeping perspective. Concepts of transparency, accountability, fairness and social justice were seen to form a common background, but it was explanation and explicability that dominated the foreground, particularly in connection to their relationship to trust.

In these terms, recordkeeping professionals should seek to concentrate on the quality of explanations that can be offered into the future. As a result of XAI research, the AI community will improve the interpretability of its techniques, but these explanations may always remain highly technical and impenetrable to those outside that community. Even so, the recordkeeping community does have frameworks, particularly in its theoretical elaboration of authenticity, that may prove useful as XAI researchers continue to elaborate and evaluate the effectiveness of the techniques and approaches they have towards interpretability.

Then again XAI, in its wider sense as a growing conversation about the accountability and trustworthiness of systems on which we increasingly rely, is already starting to envisage other forms of explanation, those which encompass documentation of the processes whereby AI techniques are both developed, deployed and applied. Recordkeeping professionals can immediately apply their expertise in documentation to assist in these efforts, addressing (if they are not already) questions such as;

- What records are created within AI research teams to document their process?
- What records are created of the decisions to procure or deploy systems utilising AI?
- What records are created of the decisions and impact of such systems?
- Are the created records sufficient to meet existing legal provisions?
- Do the created records meet the required standards of quality?

It is to be hoped that future issues of this journal will report research that explores some of these questions, for it is this way that recordkeepers can best contribute their particular expertise and perspective to the wider conversation about explainable, accountable and trustworthy artificial intelligence.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. (2018), “Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda”, in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, April 21-26, 2018*, ACM, New York, Paper No. 582.

Annual Conference on Neural Information Processing Systems (2016), “Interpretable ML for Complex Systems NIPS 2016”, available at https://sites.google.com/site/nips2016interpretml/ (accessed 20 August 2019).

Association of Computing Machinery (2017), “Statement on Algorithmic Transparency and Accountability”, available at https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf (accessed 20 August 2019.)

Bunn, J., "Workshop on Human-centered Explainable Artificial Intelligence", available at https://blogs.ucl.ac.uk/hexai/ (accessed 20 August 2019).

Bunn, J., "Participants", available at https://blogs.ucl.ac.uk/hexai/participants (accessed 20 August 2019).

Cornes, R. (1989), "Managing Information on IT", *Records Management Journal*, Vol. 1 No. 4, pp.170-172.

Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., Jagadish, H. V., Unsworth, K., Sahuguet, A., Venkatasubramanian, S., Wilson, C., Yu, C., and Zevenbergen, B. (2016), "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms", available at https://www.fatml.org/resources/principles-for-accountable-algorithms (accessed 20 August 2019).

Duff, W., Flinn, A., Suurtamm, K. E. and Wallace, D. A. (2013), "Social justice impact of archives: a preliminary investigation", *Archival Science*, Vol.13 No. 4, pp.317-348.

Gilpin, L., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018), "Explaining Explanations: An overview of interpretability of machine learning", accepted at *The 5th International Conference on Data Science and Advanced Analytics*, available at arXiv:1806.00069 [cs.AI].

Gracy, K. (2004), "Documenting Communities of Practice: Making the Case for Archival Ethnography", *Archival Science*, Vol. 4 No. 3-4, pp. 335-365.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019), *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition, available at https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/ autonomous-systems.html (accessed 28 November 2019).

Independent High-Level Expert Group on Artificial Intelligence (2019), *Ethics Guidelines for Trustworthy AI*, European Commission, Brussels, available at https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top (accessed 28 November 2019).

Information Commissioner's Office. "What else do we need to consider if Article 22 applies", available at https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling/what-else-do-we-need-to-consider-if-article-22-applies/#id3 (accessed 20 August 2019).

International Conference of Machine Learning (2016), "Workshop on Human Interpretability in Machine Learning", available at https://sites.google.com/site/2016whi/ (accessed 20 August 2019).

International Joint Conference on Artificial Intelligence (2017), "Workshop on Explainable Artificial Intelligence", available at http://home.earthlink.net/~dwaha/research/meetings/ijcai17-xai/ (accessed 20 August 2019).

International Joint Conference on Neural Networks (2017), "Special Session on Explainability of Learning Machines", available at http://gesture.chalearn.org/ijcnn17_explainability_of_learning_machines (accessed 20 August 2019).

International Standards Organization (2017), *ISOC/IEC/IEEE 12207:2017 Systems and Software Engineering – Software Life Cycle Processes*, ISO, Switzerland.InterPARES Trust (©2018), *InterPARES Trust*, available at https://interparestrust.org/ (accessed 28 November 2019).

Miller, T. (2019), "Explanation in artificial intelligence: Insights from the social sciences", *Artificial Intelligence*, Vol. 267, pp. 1-38.

Noble, S. U. (2018), *Algorithms of Oppression: How Search Engines Reinforce Racism*, NYU Press, New York.

O'Neil, C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown, New York.

Organisation for Economic Co-operation and Development (2019), "OECD Principles on AI", available at http://www.oecd.org/going-digital/ai/principles/ (accessed 28 November 2019).

Panel for the Future of Science and Technology (2019), *A governance framework for algorithmic accountability and transparency,* European Parliamentary Research Service, Scientific Foresight Unit (STOA), PE 624.262. Available at http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2019)624262

Pearce-Moses, R. (2005), *A Glossary of Archival and Records Terminology*, Society of American Archivists, Chicago. Available at https://www2.archivists.org/glossary/terms/a/accountability (accessed 20 August 2019).

Pearce-Moses, R. (ed.) (2018) *InterPARES Trust Terminology*, InterPARES Trust, available at https://interparestrust.org/terminology/term/transparency (accessed 20 August 2019).

Pfeffer, Z. (2018), "Transcript of Ali Rahimi NIPS 2017 Test-of-Time Award Presentation Speech", available at https://www.zachpfeffer.com/single-post/2018/12/04/Transcript-of-Ali-Rahimi-NIPS-2017-Test-of-Time-Award-Presentation-Speech (accessed 20 August 2019).

Rolan, G., Humphries, G., Jeffrey, L., Samaras, E., Antsoupova, T. and Stuart, K. (2019), "More human than human? Artificial intelligence in the archive", *Archives and Manuscripts*, Vol.47 No. 2, pp.179-203.

Sculley, D., Snoek, J., Rahimi, A., and Wiltschko, A. (2018), " Winner's Curse? On Pace, Progress, And Empirical Rigor", available at https://openreview.net/forum?id=rJWF0Fywf (accessed 20 August 2019).

Terminology Cross-domain Task Force (2008), “Appendix 22: InterPARES 2 Project Ontologies”, in Duranti, L. and Prestion, R. (Eds.), *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records*, Associazione Nazionale Archivistica Italiana, Padova, Italy, available at http://www.interpares.org/ip2/display_file.cfm?doc=ip2_book_appendix_22.pdf (accessed 28 November 2019).

The National Archives (2018), “Rethinking the record”, available at http://www.nationalarchives.gov.uk/about/our-research-and-academic-collaboration/our-research-and-people/our-research-priorities/rethinking-the-record/ (accessed 20 August 2019).

Turek, M. (2018), “Explainable Artificial Intelligence”, available at https://www.darpa.mil/program/explainable-artificial-intelligence (accessed 20 August 2019).