ARTICLE

Explainable Artificial Intelligence (XAI)

Adoption and Advocacy

Michael Ridley

ABSTRACT

The field of explainable artificial intelligence (XAI) advances techniques, processes, and strategies that provide explanations for the predictions, recommendations, and decisions of opaque and complex machine learning systems. Increasingly academic libraries are providing library users with systems, services, and collections created and delivered by machine learning. Academic libraries should adopt XAI as a tool set to verify and validate these resources, and advocate for public policy regarding XAI that serves libraries, the academy, and the public interest.

INTRODUCTION

Explainable artificial intelligence (XAI) is a subfield of artificial intelligence (AI) that provides explanations for the predictions, recommendations, and decisions of intelligent systems.[1] Machine learning is rapidly becoming an integral part of academic libraries. XAI is a set of techniques, processes, and strategies that libraries should adopt and advocate for to ensure that machine learning appropriately serves librarianship, the academy, and the public interest.

Knowingly or not, libraries acquire and provide access to systems, services, and collections infused and directed by machine learning methods, and library users are engaged in information behavior (e.g., seeking, using, managing) facilitated or augmented by machine learning. Machine learning in library and information science (LIS), as with many other fields, has become ubiquitous. However, this technology is often opaque and complex, yet consequential. There are significant concerns about bias, unfairness, and veracity.[2] There are troubling questions about user agency and power imbalances.[3]

While LIS has a long-standing interest in AI and intelligent information systems generally,[4] it has only recently turned its attention to XAI and how it affects the field and how the field might influence it.[5] XAI is a critical lens through which to view machine learning in libraries. It is also a set of techniques, processes, and strategies essential to influencing and shaping this still emerging technology:

> Research libraries have a unique and important opportunity to shape the development, deployment, and use of intelligent systems in a manner consistent with the values of scholarship and librarianship. The area of explainable artificial intelligence is only one component of this, but in many ways, it may be the most important.[6]

Dismissing engagement with XAI because it is "highly technical and impenetrable to those outside that community" is neither acceptable nor increasingly possible.[7] Artificial intelligence is the essential substrate of contemporary information systems and XAI is a tool set for critical assessment and accountability. The details matter and must be understood if libraries are to have a place at the table as XAI, and machine learning, evolves and further deepens its effect on LIS.

Michael Ridley (mridley@uoguelph.ca) is Librarian, University of Guelph. © 2022.

https://doi.org/10.6017/ital.v41i2.14683

This paper provides an overview of XAI with key definitions, a historical context, and examples of XAI techniques, strategies, and processes that form the basis of the field. It considers areas where XAI and academic libraries intersect. The dual emphasis is on XAI as a toolset for libraries to adopt and XAI as an area for public policy advocacy.

WHAT IS XAI?

XAI is plagued by definitional problems.[8] Some definitions are focused solely and narrowly on the technical concepts while others focus only on the broad social and political dimensions. Lacking "a theory of explainable AI, with a formal and universally agreed definition of what explanations are,"[9] the fundamentals of this field are still being explored, often from different disciplinary perspectives.[10] Critical algorithm studies position machine learning as socio-techno-informational systems.[11] As such, a definition of XAI must encompass not just the techniques, as important and necessary as they are, but also the context within which XAI operates.

The US Defense Advanced Research Projects Agency (DARPA) description of XAI captures the breadth and scope of the field. The purpose of XAI is for AI systems to have "the ability to explain their rationale, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future" [12] and to "enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners."[13] XAI is needed to:

1. generate trust, transparency, and understanding;
2. ensure compliance with regulations and legislation;
3. mitigate risk;
4. generate accountable, reliable, and sound models for justification;
5. minimize or mitigate bias, unfairness, and misinterpretation in model performance and interpretation; and
6. validate models and validate explanations generated by XAI.[14]

XAI consists of testable and unambiguous proofs, various verification and validation methods that assess influence and veracity, and authorizations that define requirements or mandate auditing within a public policy framework.

XAI is not a new consideration. Explainability has been a preoccupation of computer science since the early days of expert systems in the late twentieth century.[15] However, the 2018 introduction of the General Data Protection Regulation (GDPR) by the European Union (EU) shifted explainability from a purely technical issue to one with an additional and urgent focus on public policy.[16] While the presence of a "right to explanation" in the GDPR is highly contested, [17] industry groups and jurisdictions beyond the EU recognized its evitability spurring an explosion in XAI research and development.[18]

TYPES OF XAI

Taxonomies of XAI types are classified based on their scope and mechanism.[19] Local explanations interpret the decisions of a machine learning model used in a specific instance (i.e., involving data and context relevant to the circumstance). Global explanations interpret the model more generally (i.e., involving all the training data and relevant contexts). In black-box or model-agnostic explanations, only the input and the output of the machine learning model are required while white-box or model-specific explanations require more detailed information regarding the processing or design of the model.

Another way to categorize XAI is as proofs, validations, and authorizations. Proofs are testable, traceable, and unambiguous explanations demonstrable through causal links, logic statements, or transparent processes. Typically, proofs are only available for AI systems that use "inherently interpretable" techniques such as rules, decisions trees, or linear regressions.[20]

Validations are explanations that confirm the veracity of the AI system. These verifications occur through testing procedures, reproducibility, approximations and abstractions, and justifications.

Authorizations are explanations because of processes in which third parties provide some form of standard, ratification, prohibition, or audit. Authorizations might pertain to the AI model, its operation in specific instances, or even the process by which the AI was created. They can be provided by professional groups, nongovernmental organizations, governments and government agencies, and third parties in the public and private sector.

Academic libraries can adopt proofs and validations as means to interrogate information systems and resources. This includes collections which are increasingly machine learning systems themselves or developed with machine learning methods. The recognition of "collections as data" is an important shift in this direction.[21] Where appropriate, proofs and validations should accompany content and systems derived from machine learning. Libraries must also engage with XAI as authorizations to assess the public policy implications that exist, are emergent, or are necessary. Library advocacy is currently lacking in this area. The requirement for policy and governance frameworks is a reminder that machine learning is "far from being purely mechanistic, it is deeply, inescapably human"[22] and that while complex and opaque "the 'black box' is full of people."[23]

PREREQUISITES TO AN XAI STRATEGY

Three questions are important for any XAI strategy:

- What constitutes a good explanation?
- Who is the explanation for?
- How will the explanation be provided?

Explanations are context specific. The "goodness" of an explanation is dependent on the needs and objectives of the explainee (a user) and the explainer (an XAI). Following research from the fields of psychology and cognitive science, Keil suggests five reasons for why someone wants an explanation: (1) to predict similar events in the future, (2) to diagnose, (3) to assess blame or guilt, (4) to justify or rationalize an action, and (5) for aesthetic pleasure.[24]

For most people, explanations need not be complete or even fully accurate.[25] As a result, who the explanation is for is critical to a good explanation. Different audiences have different priorities. System developers are primarily interested in performance explanations while clients focus on effectiveness or efficacy, professionals are concerned about veracity, and regulators are interested in policy implications. Nonexpert, lay users of a system want explanations that build trust and provide accountability.

A good explanation is also affected by its presentation. There are temporal and format considerations. Explanations can be provided or available in real time and continuously as the process occurs (hence partial explanations) or post hoc and in summary form. Interactive explanations are widely preferred but are not always appropriate or actionable.[26] Studies have compared textual, visual, and multimodal formats with differing results. Familiar textual responses or simple visual explanations such as Venn diagrams are often most effective for nonexpert users.[27]

Drawing from philosophy, psychology, and cognitive science, Miller recommends four approaches for XAI.[28] Explanations are contrastive. When people want to know the "why" of something, "people do not ask why event P happened, but rather why event P happened *instead* of some event Q." Explanations are selected. "Humans are adept at selecting one or two causes from a sometimes infinite number of causes to be *the* explanation." Explanations are social. "They are a transfer of knowledge, presented as part of a conversation or interaction, and are thus presented relative to the explainer's beliefs about the explainee's beliefs." Finally, Miller cautions against using probabilities and statistical relationships and encourages references to causes.

Burrell identifies three key barriers to explainability: concealment, the limited technical understanding of the user, and an incompatibility between the user (human) and algorithmic reasoning.[29] While concealment is deliberate, it may or may not be justified. Protecting IP and trade secrets is acceptable while obscuring processes to purposively deceive users is not. Regulations are a tool to moderate the former and minimize the latter.

The technical limitations of users and the incompatibility between users and algorithms suggest two remedies. First is enhancing algorithmic literacy. Algorithmic literacy is a "a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace."[30] Libraries have a key role in advancing algorithmic literacy in their communities.[31] Just as libraries championed information literacy through the promulgation of standards and principles, the provision of diverse educational programming, and the engagement of the broad academic community, so too can libraries be central to efforts to enhance algorithmic literacy. Second is a requirement that XAI must be sensitive to the abilities and needs of different users. A survey of the key challenges and research direction of XAI identified 39 issues, including the need to understand and enhance the user experience, match XAI to user expertise, and explain the competencies of AI systems to users.[32] This is the essence of human-centered explainable AI (HCXAI). Among HCXAI principles are the importance of context (regarding user objectives, decision consequences, timing, modality, and intended audience), the value of using hybrid explanation methods that complement and extend each other, and the power of contrastive examples and approaches.[33]

PROOFS AND VALIDATIONS

XAI that provide proofs or validations can be adopted by libraries to assess and evaluate machine learning utilized in systems, services, and collections. Since proofs pertain to already interpretable systems, the four examples provided focus on validations: feature audit, approximation and abstraction, reproducibility, and XAI by AI.

These techniques may require access to, or information about, the machine learning model. This would include such characteristics as the algorithms used, settings of the parameters and hyperparameters, optimization choices, and the training data. While all these may not be normally available, designers of machine learning systems in consequential settings should expect to provide, indeed be required to provide, such access. Similarly, vendors of library content or systems utilizing machine learning should make explanatory proofs and validations available for library inspection.

Feature Audit

Feature audit is an explanatory strategy that attempts to reveal the key features (e.g., characteristics of the data or settings of the hyperparameters used to the differentiate data) that have a primary role in the prediction of the algorithm. By isolating these features, it is possible to explain the key components of the decision. Feature audit is a standard technique of linear regression, but it is made more difficult in machine learning because of the complexity of the information space (e.g., billions of parameters and high dimensionality). There are various feature audit techniques[34] but all of them are "decompositional" in that they attempt to reduce the work of the algorithm to its component parts and then use those results as an explanation.[35] Feature audit can highlight bias or inaccuracy by revealing incongruence between the data and the prediction. More advanced feature audit techniques (e.g., gradient feature auditing) recognize that features can indirectly influence other features and that these features are not easily detectable as separate, influential elements.[36] This interaction among features challenges the strict decompositional approach to feature audit and will likely lead to an increased focus on the relational analysis among and between elements.

Approximation and Abstraction

Approximation and abstraction are techniques that create a more simplified model to explain the more complex model.[37] People seek and accept explanations that "satisfice"[38] and are coherent with existing beliefs.[39] This recognizes that "an explanation has greater power than an alternative if it makes what is being explained less surprising."[40]

Approaches such as "model distillation"[41] or the "model agnostic" feature reduction of the Local Interpretable Model-Agnostic Explanations (LIME) tool create a simplified presentation of the algorithmic model.[42] This approximation or abstraction may compromise accuracy, but it provides an accessible representation that enhances understandability.

A different type of approximation or abstraction is a narrative of the machine learning processes utilized that provides sufficient documentation for a reader to act as an explanation of the outcomes. An exemplary case of this is *Lithium-Ion Batteries: A Machine-Generated Summary of Current Research* published by Springer Nature and written by Beta Writer, an AI or more accurately a suite of algorithms.[43] A collaboration of machine learning and human editors, the full production cycle of the book is documented in the introduction.[44] In lieu of being able to interrogate the system directly, this detailed account provides an explanation of the system allowing readers to assess the strengths, limitations, and confidence levels of the algorithmic processes and offers a model of what might be necessary for future AI generated texts.[45] Libraries can utilize this documentation in acquisition or licensing decisions and subsequently make it available as user guides when resources are added to the collection.

Reproducibility

Replication is a verification strategy fundamental to science. Being able to independently reproduce results in different settings provides evidence of veracity and supports user trust. However, documented problems in reproducing machine learning studies have questioned the generalizability of these approaches and undermined their explanatory capacity. For example, an analysis of text mining studies using machine learning for citation screening in the preparation of systemic reviews revealed a lack of key elements to enable replicability (e.g., access to research datasets, software environments used, randomization control, and lack of detail on new methods proposed or employed).[46] In response, a "Reproducibility Challenge" was created by the International Conference on Learning Representations (ICLR) to validate 2018 conference submissions and has continued in subsequent meetings.[47] More rigorous replication through the availability of all necessary components and the development of standards will be important to this type of verification.[48]

XAI by AI

The inherent complexity and opacity of unsupervised learning or reinforcement learning suggests, as XAI researcher Trevor Darrell puts it, "the solution to explainable AI is more AI."[49] In this approach to explanation, oversight AI are positioned as intermediaries between an AI and its users:

> Workers have supervisors; businesses have accountants; schoolteachers have principals. We suggest that the time has come to develop AI oversight systems ("AI Guardians") that will seek to ensure that the various smart machines will not stray from the guidelines their programmers have provided.[50]

While the prospect of AI guardians may be dystopic, oversight systems performing roles that validate, interrogate, and report are common in code checking tools. Generative adversarial networks (GANs) have been used to create counterfactual explanations of another machine learning model to enhance explainability.[51] With strategic organizational and staffing changes to enhance capabilities, libraries can design and deploy such oversight or adversarial tools with objectives appropriate to the requirements and norms of libraries and the academy.

AUTHORIZATION

XAI that results from authorizations is an area where public policy engagement is needed to ensure XAI, and machine learning, are appropriately serving libraries, the academy, and the public at large. Three examples are provided: codes and standards, regulation, and audit.

Codes and Standards

One approach to explanation, supported by the AI industry and professional organizations, are voluntary codes or standards that encourage explanatory capabilities. These nonbinding principles are a type of self-regulation and are widely promoted as a means of assurance.[52]

The Association for Computing Machinery's statement on algorithms highlights seven principles as guides to system design and use: awareness, access and redress, accountability, explanation, data provenance, auditability, validation, and testing. However, the language used is tentative and conditional. Designers are "encouraged" to provide explanations and to "encourage" a means for interrogation and auditing "where harm is suspected" (i.e., a post hoc process). Despite this, the statement concludes with a strong position on accountability if not explainability: "Institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results."[53]

Unfortunately, the optimism for self-regulation in explainability is undercut by the poor experience with voluntary mechanisms regarding privacy protection.[54] In addition, library associations, library system vendors, and scholarly publishers have been slow to endorse any codes or standards regarding explainability.

Regulation

The most common recommendation for AI oversight and authorization to ensure explainability is the creation of a regulatory agency. Specific suggestions include a "neutral data arbiter" with investigative powers like the US Federal Trade Commission,[55] a Food and Drug Administration "for algorithms,"[56] a standing "Commission on Artificial Intelligence,"[57] quasi-governmental agencies such as the Council of Europe,[58] and a hybrid agency model combining certification and liability.[59] Such agencies would have legislated or delegated powers to investigate, certify, license, and arbitrate on matters relating to AI and algorithms, including their design, use, and effects. There are few calls for an international regulatory agency despite digitally porous national boundaries and the global reach of machine learning.[60]

That almost no such agencies have been created reveals the strength and influence of the large corporations responsible for developing and deploying most machine learning tools and systems.[61] Reports comparing regulatory approaches to AI among the European Union, the United Kingdom, the United States, and Canada indicate significantly different approaches but with most proceeding with a "light touch" to avoid competitive disadvantages in a multitrillion dollar global marketplace.[62]

The introduction of the draft EU Artificial Intelligence Act marks the first major jurisdiction to propose specific AI legislation.[63] While the act is fulsome about high-risk AI, it is silent on any notion of "explainable" AI, preferring to focus on the less specific idea of "trustworthy artificial intelligence." With this the EU appears to retreat from the idea of explainability in the GDPR.

An exception to this inertia or backtracking is the development and use of algorithmic impact assessments in both governments and industry. These instruments help prospective users of an algorithmic decision-making system determine levels of explanatory requirements and standards to meet those requirements.[64] Canada has been a leader in this area with a protocol covering use of these systems in the federal government.[65]

Some identify due process as a possible, if limited, remedy for explainability.[66] However, a landmark US case suggests otherwise. In *State v. Loomis*, regarding the use of COMPAS, an algorithmic sentencing system, the court ruled on the role of explanation in due process:[67]

> The Wisconsin Supreme Court held that a trial court's use of an algorithmic risk assessment in sentencing did not violate the defendant's due process rights even though the methodology used to produce the assessment was disclosed neither to the court nor to the defendant.[68]

The petition of the Loomis case to the US Supreme Court was denied, so a higher court ruling on this issue is unavailable.[69]

Advocacy for regulations regarding explainability should be a central concern for libraries. Without strong regulatory oversight requiring disclosure and accountability, machine learning systems will remain black boxes and presence of these consequential systems in the lives of users will be obscured.

Audit

A commonly recommended approach to AI oversight and explanation is third-party auditing.[70] The use of audit and principles of auditing are widely accepted in a variety of areas.[71] In a library context, auditing of AI can be thought of as a reviewing process to achieve transparency or to determine product compliance. Auditing is typically done after system implementation, but it can be accomplished at any stage. It is possible to audit design specifications, completed code, cognitive models, or periodic audits of specific decisions.[72] The keys to successful audit oversight are clear audit goals and objectives (e.g., what is being audited and for what purpose), acknowledged expertise of the auditors, authority of the auditors to recommend, and authorization of the auditors to investigate. Any such auditing responsibility for XAI would require the trust of stakeholders such as AI designers, government regulators, industry representatives as well as users themselves.

Critics of the audit approach have focused on lack of auditor expertise, algorithmic complexity, and the need for approaches that assess the algorithmic system prior to its release.[73] While most audit recommendations assume a public agency in this role, an innovative suggestion is a crowdsourced audit (a form of audit study that involves the recruitment of testers to anonymously assess an algorithmic system; an XAI form of the "secret shopper").[74] This approach resembles techniques used by consumer advocates and might indicate the rise of public activists into the XAI arena.

The complexity of algorithms suggests that a precondition for an audit is "auditability."[75] This would require that AI be designed in such a way that an audit is possible (i.e., inspectable in some manner) while, presumably, not impairing its predictive performance. Sandvig et al. propose regulatory changes because "rather than regulating for transparency or misbehavior, we find this situation argues for 'regulation toward auditability'."[76]

Auditing is not without its difficulties. There are no industry standards for algorithmic auditing.[77] A high-profile development was the recent launch of ORCAA (orcaarisk.com), an algorithmic auditing company started by Cathy O'Neil, a data scientist who has written extensively about the perils of uncontrolled algorithms.[78] However, the legitimacy of third-party auditing has been criticized as lacking public transparency and the capacity to demand change.[79]

While libraries may not be able to create their own auditing capacity, whether collectively or individually, they are encouraged to engage with the emerging algorithmic auditing community to shape auditing practices appropriate for scholarly communication.

XAI AS DISCOVERY

While XAI is primarily a means to validate and authorize machine learning systems, another use of XAI is gaining attention. Since XAI can find new information latent in large and complex datasets, discovery is promoted as "one of the most important achievements of the entire algorithmic explainability project."[80] Alkhateeb asks "can scientific discovery really be automated" while invoking the earlier work of Swanson which mined the medical literature for new knowledge by connecting seemingly unrelated articles through search.[81] An emerging reason for libraries to adopt XAI may be as a powerful discovery tool.

CONCLUSION

Our lives have become "algorithmically mediated"[82] where we are "dependent on computational spectacles to see the world."[83] Academic libraries are now sites where systems, services, and collections are increasingly shaped and provided by machine learning. The predictions, recommendations, and decisions of machine learning systems are powerful as well as consequential. However, "the danger is not so much in delegating cognitive tasks, but in distancing ourselves from—or in not knowing about—the nature and precise mechanisms of that delegation."[84] Taddeo notes that "delegation without supervision characterises the presence of trust."[85] XAI is an essential tool to build that trust.

Geoffrey Hinton, a central figure in the development of machine learning,[86] argues that requiring an explanation from an AI system would be "a complete disaster" and that trust and acceptance should be based on the system's performance, not its explainability.[87] This is consistent with the view of many that "if algorithms that cannot be easily explained consistently make better decisions in certain areas, then policymakers should not require an explanation."[88] Both these views are at odds with the tenants of critical thought and assessment, and both challenge norms of algorithmic accountability.

XAI is a dual opportunity for libraries. On one hand, it is a set of techniques, processes, and strategies that enable the interrogation of the algorithmically driven resources that libraries provide to their users. On the other hand, it is a public policy arena where advocacy is necessary to promote and uphold the values of librarianship, the academy, and the public interest in the face of powerful new technologies. Many disciplines have engaged with XAI as machine learning has impacted their fields.[89] XAI has been called a "disruptive force" in LIS,[90] warranting the growing interest in how XAI affects the field and how the field might influence it.

ENDNOTES

[1] Vijay Arya et al., “One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques,” *ArXiv:1909.03012 [Cs, Stat]*, 2019, http://arxiv.org/abs/1909.03012; Shane T. Mueller et al., “Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI,” *ArXiv:1902.01876 [Cs]*, 2019, http://arxiv.org/abs/1902.01876; Ingrid Nunes and Dietmar Jannach, “A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems,” *User Modeling and User-Adapted Interaction* 27, no. 3 (2017): 393–444, https://doi.org/10.1007/s11257-017-9195-0; Gesina Schwalbe and Bettina Finzel, “XAI Method Properties: A (Meta-) Study,” *ArXiv:2105.07190 [Cs]*, 2021, http://arxiv.org/abs/2105.07190.

[2] Safiya Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: New York University Press, 2018); Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, Mass.: Harvard University Press, 2015); Sara Wachter-Boettcher, *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech* (New York: W. W. Norton, 2017).

[3] Abeba Birhane et al., “The Values Encoded in Machine Learning Research,” *ArXiv:2106.15590 [Cs]*, 2021, http://arxiv.org/abs/2106.15590; Taina Bucher, *If ... Then: Algorithmic Power and Politics* (New York: Oxford University Press, 2018); Sarah Myers West, Meredith Whittaker, and Kate Crawford, Discriminating Systems: Gender, Race, and Power in AI (AI Now Institute, 2019), https://ainowinstitute.org/discriminatingsystems.html.

[4] Rao Aluri and Donald E. Riggs, “Application of Expert Systems to Libraries,” ed. Joe A. Hewitt, *Advances in Library Automation and Networking* 2 (1988): 1–43; Ryan Cordell, *Machine Learning + Libraries: A Report on the State of the Field* (Washington DC: Library of Congress, 2020), https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf; Jason Griffey, ed., “Artificial Intelligence and Machine Learning in Libraries,” *Library Technology Reports* 55, no. 1 (2019), https://doi.org/10.5860/ltr.55n1; Guoying Liu, “The Application of Intelligent Agents in Libraries: A Survey,” *Program: Electronic Library and Information Systems* 45, no. 1 (2011): 78–97, https://doi.org/10.1108/00330331111107411; Linda C. Smith, “Artificial Intelligence in Information Retrieval Systems,” *Information Processing and Management* 12, no. 3 (1976): 189–222, https://doi.org/10.1016/0306-4573(76)90005-4.

[5] Jenny Bunn, “Working in Contexts for Which Transparency Is Important: A Recordkeeping View of Explainable Artificial Intelligence (XAI),” *Records Management Journal (London, England)* 30, no. 2 (2020): 143–53, https://doi.org/10.1108/RMJ-08-2019-0038; Cordell, “Machine Learning + Libraries”; Andrew M. Cox, *The Impact of AI, Machine Learning, Automation and Robotics on the Information Professions* (CILIP, 2021), http://www.cilip.org.uk/resource/resmgr/cilip/research/tech_review/cilip_–_ai_report_-_final_lo.pdf; Daniel Johnson, *Machine Learning, Libraries, and Cross-Disciplinary Research: Possibilities and Provocations* (Notre Dame, Indiana: Hesburgh Libraries, University of Notre Dame, 2020), https://dx.doi.org/10.7274/r0-wxg0-pe06; Sarah Lippincott, *Mapping the Current Landscape of Research Library Engagement with Emerging Technologies in Research and Learning* (Washington DC: Association of Research Libraries, 2020), https://www.arl.org/wp-content/uploads/2020/03/2020.03.25-emerging-technologies-

landscape-summary.pdf; Thomas Padilla, *Responsible Operations. Data Science, Machine Learning, and AI in Libraries* (Dublin, OH: OCLC Research, 2019), https://doi.org/10.25333/xk7z-9g97; Michael Ridley, "Explainable Artificial Intelligence," *Research Library Issues*, no. 299 (2019): 28–46, https://doi.org/10.29242/rli.299.3.

[6] Ridley, "Explainable Artificial Intelligence," 42.

[7] Bunn, "Working in Contexts for Which Transparency Is Important," 151.

[8] Sebastian Palacio et al., "XAI Handbook: Towards a Unified Framework for Explainable AI," *ArXiv:2105.06677 [Cs]*, 2021, http://arxiv.org/abs/2105.06677; Sahil Verma et al., "Pitfalls of Explainable ML: An Industry Perspective," in *MLSYS JOURNE Workshop*, 2021, http://arxiv.org/abs/2106.07758; Giulia Vilone and Luca Longo, "Explainable Artificial Intelligence: A Systematic Review," *ArXiv:2006.00093 [Cs]*, 2020, http://arxiv.org/abs/2006.00093.

[9] Wojciech Samek and Klaus-Robert Muller, "Towards Explainable Artificial Intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ed. Wojciech Samek et al., Lecture Notes in Artificial Intelligence 11700 (Cham: Springer International Publishing, 2019), 17.

[10] Mueller et al., "Explanation in Human-AI Systems."

[11] Isto Huvila et al., "Information Behavior and Practices Research Informing Information Systems Design," *Journal of the Association for Information Science and Technology*, 2021, 1–15, https://doi.org/10.1002/asi.24611.

[12] DARPA, *Explainable Artificial Intelligence (XAI)* (Arlington, VA: DARPA, 2016), http://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf.

[13] Matt Turek, "Explainable Artificial Intelligence (XAI)," DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence.

[14] Julie Gerlings, Arisa Shollo, and Ioanna Constantiou, "Reviewing the Need for Explainable Artificial Intelligence (XAI)," in *Proceedings of the Hawaii International Conference on System Sciences*, 2020, http://arxiv.org/abs/2012.01007.

[15] William J. Clancey, "The Epistemology of a Rule-Based Expert System—a Framework for Explanation," *Artificial Intelligence* 20, no. 3 (1983): 215–51, https://doi.org/10.1016/0004-3702(83)90008-5; William Swartout, "XPLAIN: A System for Creating and Explaining Expert Consulting Programs," *Artificial Intelligence* 21 (1983): 285–325; William Swartout, Cecile Paris, and Johanna Moore, "Design for Explainable Expert Systems," *IEEE Expert-Intelligent Systems & Their Applications* 6, no. 3 (1991): 58–64, https://doi.org/10.1109/64.87686.

[16] European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016," 2016, http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679.

[17] Lilian Edwards and Michael Veale, "Slave to the Algorithm? Why a 'Right to Explanation' Is Probably Not the Remedy You Are Looking For," *Duke Law & Technology Review* 16 (2017): 18–84; Bryce Goodman and Seth Flaxman, "European Union Regulations on Algorithmic Decision Making and a 'Right to Explanation'," *AI Magazine* 38, no. 3 (2017): 50–57, https://doi.org/10.1609/aimag.v38i3.2741; Margot E. Kaminski, "The Right to Explanation, Explained," *Berkeley Technology Law Journal* 34, no. 1 (2019): 189–218, https://doi.org/10.15779/Z38TD9N83H; Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *International Data Privacy Law* 7, no. 2 (2017): 76–99, https://doi.org/10.1093/idpl/ipx005.

[18] Amina Adadi and Mohammed Berrada, "Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access* 6 (2018): 52138–60, https://doi.org/10.1109/ACCESS.2018.2870052; Mueller et al., "Explanation in Human-AI Systems"; Vilone and Longo, "Explainable Artificial Intelligence."

[19] Schwalbe and Finzel, "XAI Method Properties."

[20] Or Biran and Courtenay Cotton, "Explanation and Justification in Machine Learning: A Survey" (International Joint Conference on Artificial Intelligence, workshop on Explainable Artificial Intelligence (XAI), Melbourne, 2017), http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf.

[21] Padilla, *Responsible Operations*.

[22] Jenna Burrell and Marion Fourcade, "The Society of Algorithms," *Annual Review of Sociology* 47, no. 1 (2021): 231, https://doi.org/10.1146/annurev-soc-090820-020800.

[23] Nick Seaver, "Seeing like an Infrastructure: Avidity and Difference in Algorithmic Recommendation," *Cultural Studies* 35, no. 4–5 (2021): 775, https://doi.org/10.1080/09502386.2021.1895248.

[24] Frank C. Keil, "Explanation and Understanding," *Annual Review of Psychology* 57 (2006): 227–54, https://doi.org/10.1146/annurev.psych.57.102904.190100.

[25] Donald A. Norman, "Some Observations on Mental Models," in *Mental Models*, ed. Dedre Gentner and Albert L. Stevens (New York: Psychology Press, 1983), 7–14.

[26] Ashraf Abdul et al., "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*, CHI '18 (New York: ACM, 2018), 582:1–582:18, https://doi.org/10.1145/3173574.3174156; Joachim Diederich, "Methods for the Explanation of Machine Learning Processes and Results for Non-Experts," *PsyArXiv*, 2018, https://doi.org/10.31234/osf.io/54eub.

[27] Pigi Kouki et al., "User Preferences for Hybrid Explanations," in *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17 (New York, NY: ACM, 2017), 84–88, https://doi.org/10.1145/3109859.3109915.

[28] Tim Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial Intelligence* 267 (2019): 3, https://doi.org/10.1016/j.artint.2018.07.007.

[29] Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms," *Big Data & Society* 3, no. 1 (2016), https://doi.org/10.1177/2053951715622512.

[30] Duri Long and Brian Magerko, "What Is AI Literacy? Competencies and Design Considerations," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20 (Honolulu, HI: Association for Computing Machinery, 2020), 2, https://doi.org/10.1145/3313831.3376727.

[31] Michael Ridley and Danica Pawlick-Potts, "Algorithmic Literacy and the Role for Libraries," *Information Technology and Libraries* 40, no. 2 (2021), https://doi.org/doi.org/10.6017/ital.v40i2.12963.

[32] Waddah Saeed and Christian Omlin, "Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities," *ArXiv:2111.06420 [Cs]*, 2021, http://arxiv.org/abs/2111.06420.

[33] Shane T. Mueller et al., "Principles of Explanation in Human-AI Systems" (Explainable Agency in Artificial Intelligence Workshop, AAAI 2021), http://arxiv.org/abs/2102.04972.

[34] Sebastian Bach et al., "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLoS ONE* 10, no. 7 (2015): e0130140, https://doi.org/10.1371/journal.pone.0130140; Biran and Cotton, "Explanation and Justification in Machine Learning: A Survey"; Chris Brinton, "A Framework for Explanation of Machine Learning Decisions" (IJCAI-17 Workshop on Explainable AI (XAI), Melbourne: IJCAI, 2017), http://www.intelligentrobots.org/files/IJCAI2017/IJCAI-17_XAI_WS_Proceedings.pdf; Chris Olah, Alexander Mordvintsev, and Ludwig Schubert, "Feature Visualization," *Distill*, November 7, 2017, https://doi.org/10.23915/distill.00007.

[35] Edwards and Veale, "Slave to the Algorithm?"

[36] Philip Adler et al., "Auditing Black-Box Models for Indirect Influence," *Knowledge and Information Systems* 54 (2018): 95–122, https://doi.org/10.1007/s10115-017-1116-3.

[37] Alisa Bokulich, "How Scientific Models Can Explain," *Synthese* 180, no. 1 (2011): 33–45, https://doi.org/10.1007/s11229-009-9565-1; Keil, "Explanation and Understanding."

[38] Herbert A. Simon, "What Is an 'Explanation' of Behavior?," *Psychological Science* 3, no. 3 (1992): 150–61, https://doi.org/10.1111/j.1467-9280.1992.tb00017.x.

[39] Norbert Schwarz et al., "Ease of Retrieval as Information: Another Look at the Availability Heuristic," *Journal of Personality and Social Psychology* 61, no. 2 (1991): 195–202, https://doi.org/10.1037/0022-3514.61.2.195; Paul Thagard, "Evaluating Explanations in Law, Science, and Everyday Life," *Current Directions in Psychological Science* 15, no. 3 (2006): 141–45, https://doi.org/10.1111/j.0963-7214.2006.00424.x.

[40] Tania Lombrozo, "Explanatory Preferences Shape Learning and Inference," *Trends in Cognitive Sciences* 20, no. 10 (2016): 756, https://doi.org/10.1016/j.tics.2016.08.001.

[41] Sarah Tan et al., "Detecting Bias in Black-Box Models Using Transparent Model Distillation," *ArXiv:1710.06169 [Cs, Stat]*, November 18, 2017, http://arxiv.org/abs/1710.06169.

[42] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Model-Agnostic Interpretability of Machine Learning," *ArXiv:1606.05386 [Cs, Stat]*, 2016, http://arxiv.org/abs/1606.05386.

[43] Beta Writer, *Lithium-Ion Batteries: A Machine-Generated Summary of Current Research* (Heidelberg: Springer Nature, 2019), https://link.springer.com/book/10.1007/978-3-030-16800-1.

[44] Henning Schoenenberger, Christian Chiarcos, and Niko Schenk, preface to *Lithium-Ion Batteries; A Machine-Generated Summary of Current Research,* by Beta Writer, (Heidelberg: Springer International Publishing, 2019).

[45] Michael Ridley, "Machine Information Behaviour," in *The Rise of AI: Implications and Applications of Artificial Intelligence in Academic Libraries*, ed. Sandy Hervieux and Amanda Wheatley (Association of College and University Libraries, 2022).

[46] Babatunde Kazeem Olorisade, Pearl Brereton, and Peter Andras, "Reproducibility of Studies on Text Mining for Citation Screening in Systematic Reviews: Evaluation and Checklist," *Journal of Biomedical Informatics* 73 (2017): 1–13, https://doi.org/10.1016/j.jbi.2017.07.010; Babatunde K. Olorisade, Pearl Brereton, and Peter Andras, "Reproducibility in Machine Learning-Based Studies: An Example of Text Mining," in *Reproducibility in ML Workshop* (International Conference on Machine Learning, Sydney, Australia, 2017), https://openreview.net/pdf?id=By4l2PbQ-.

[47] Joelle Pineau, "Reproducibility Challenge," October 6, 2017, http://www.cs.mcgill.ca/~jpineau/ICLR2018-ReproducibilityChallenge.html.

[48] Benjamin Haibe-Kains et al., "Transparency and Reproducibility in Artificial Intelligence," *Nature* 586, no. 7829 (2020): E14–E16, https://doi.org/10.1038/s41586-020-2766-y; Benjamin J. Heil et al., "Reproducibility Standards for Machine Learning in the Life Sciences," *Nature Methods*, August 30, 2021, https://doi.org/10.1038/s41592-021-01256-7.

[49] Cliff Kuang, "Can A.I. Be Taught to Explain Itself?," *The New York Times Magazine*, November 21, 2017, 50, https://nyti.ms/2hR1S15.

[50] Amitai Etzioni and Oren Etzioni, "Incorporating Ethics into Artificial Intelligence," *The Journal of Ethics* 21, no. 4 (2017): 403–18, https://doi.org/10.1007/s10892-017-9252-2.

[51] Kamran Alipour et al., "Improving Users' Mental Model with Attention-Directed Counterfactual Edits," *Applied AI Letters*, 2021, e47, https://doi.org/10.1002/ail2.47.

[52] Association for Computing Machinery, *Statement on Algorithmic Transparency and Accountability* (New York: ACM, 2017), http://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf; Alex Campolo et al., *AI Now 2017 Report* (New

York: AI Now Institute, 2017); IEEE, *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems* (New York: IEEE, 2019), https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf.

[53] Association for Computing Machinery, *Statement on Algorithmic Transparency and Accountability*, 2.

[54] Lilian Edwards and Michael Veale, "Enslaving the Algorithm: From a 'Right to an Explanation' to a 'Right to Better Decisions'?," *IEEE Security & Privacy* 16, no. 3 (2018): 46–54.

[55] Kate Crawford and Jason Schultz, "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms," *Boston College Law Review* 55, no. 1 (2014): 93–128.

[56] Andrew Tutt, "An FDA for Algorithms," *Administrative Law Review* 69, no. 1 (2017): 83–123.

[57] Corinne Cath et al., "Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach," *Science and Engineering Ethics*, March 28, 2017, https://doi.org/10.1007/s11948-017-9901-7.

[58] Edwards and Veale, "Slave to the Algorithm?"

[59] Matthew U. Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies," *Harvard Journal of Law & Technology* 29, no. 2 (2016): 353–400.

[60] Roger Brownsword, "From Erewhon to AlphaGo: For the Sake of Human Dignity, Should We Destroy the Machines?," *Law, Innovation and Technology* 9, no. 1 (January 2, 2017): 117–53, https://doi.org/10.1080/17579961.2017.1303927.

[61] Birhane et al., "The Values Encoded in Machine Learning Research"; Ana Brandusescu, *Artificial Intelligence Policy and Funding in Canada: Public Investments, Private Interests* (Montreal: Centre for Interdisciplinary Research on Montreal, McGill University, 2021).

[62] Cath et al., "Artificial Intelligence and the 'Good Society'"; Law Commission of Ontario and Céline Castets-Renard, *Comparing European and Canadian AI Regulation*, 2021, https://www.lco-cdo.org/wp-content/uploads/2021/12/Comparing-European-and-Canadian-AI-Regulation-Final-November-2021.pdf.

[63] European Commission, "Artificial Intelligence Act," 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[64] Dillon Reisman et al., *Algorithmic Impact Assessment: A Practical Framework for Public Agency Accountability* (New York: AI Now Institute, 2018), https://ainowinstitute.org/aiareport2018.pdf.

[65] Treasury Board of Canada Secretariat, "Directive on Automated Decision-Making," 2019, http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

[66] Danielle Keats Citron and Frank Pasquale, “The Scored Society: Due Process for Automated Predictions,” *Washington Law Review* 89 (2014): 1–33; Scherer, “Regulating Artificial Intelligence Systems.”

[67] Julia Angwin et al., “Machine Bias,” *ProPublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[68] “State v. Loomis,” *Harvard Law Review* 130, no. 5 (2017), https://harvardlawreview.org/2017/03/state-v-loomis/.

[69] “Loomis v. Wisconsin,” SCOTUSblog, June 26, 2017, http://www.scotusblog.com/case-files/cases/loomis-v-wisconsin/.

[70] Brownsword, “From Erewhon to AlphaGo”; Campolo et al., *AI Now 2017 Report*; IEEE, *Ethically Aligned Design*; Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*; Wachter, Mittelstadt, and Floridi, “Why a Right to Explanation.”

[71] Michael Power, *The Audit Society: Rituals of Verification* (Oxford: Oxford University Press, 1997).

[72] Alfred Ng, “Can Auditing Eliminate Bias from Algorithms?,” *The Markup*, February 23, 2021, https://themarkup.org/ask-the-markup/2021/02/23/can-auditing-eliminate-bias-from-algorithms.

[73] Joshua Alexander Knoll, “Accountable Algorithms” (PhD diss, Princeton University, 2015).

[74] Christian Sandvig et al., “Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms,” *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, 2014, http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.

[75] Association for Computing Machinery, *Statement on Algorithmic Transparency and Accountability*.

[76] Sandvig et al., “Auditing Algorithms,” 17.

[77] Ng, “Can Auditing Eliminate Bias from Algorithms?”

[78] Cathy O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Crown, 2016).

[79] Emanuel Moss et al., *Assembling Accountability: Algorithmic Impact Assessment for the Public Interest* (Data & Society, 2021), https://datasociety.net/wp-content/uploads/2021/06/Assembling-Accountability.pdf.

[80] David S. Watson and Luciano Floridi, “The Explanation Game: A Formal Framework for Interpretable Machine Learning,” *Synthese (Dordrecht)* 198, no. 10 (2020): 9214, https://doi.org/10.1007/s11229-020-02629-9.

[81] Ahmed Alkhateeb, "Science Has Outgrown the Human Mind and Its Limited Capacities," Aeon, April 24, 2017, https://aeon.co/ideas/science-has-outgrown-the-human-mind-and-its-limited-capacities; Don R. Swanson, "Undiscovered Public Knowledge," *The Library Quarterly* 56, no. 2 (1986): 103–18; Don R. Swanson, "Medical Literature as a Potential Source of New Knowledge.," *Bulletin of the Medical Library Association* 78, no. 1 (1990): 29–37.

[82] Jack Anderson, "Understanding and Interpreting Algorithms: Toward a Hermeneutics of Algorithms," *Media, Culture & Society* 42, no. 7–8 (2020): 1479–94, https://doi.org/10.1177/0163443720919373.

[83] Ed Finn, "Algorithm of the Enlightenment," *Issues in Science and Technology* 33, no. 3 (2017): 24.

[84] Jos de Mul and Bibi van den Berg, "Remote Control: Human Autonomy in the Age of Computer-Mediated Agency," in *Law, Human Agency, and Autonomic Computing*, ed. Mireille Hildebrandt and Antoinette Rouvroy (Abingdon: Routledge, 2011), 59.

[85] Mariarosaria Taddeo, "Trusting Digital Technologies Correctly," *Minds and Machines* 27, no. 4 (2017): 565, https://doi.org/10.1007/s11023-017-9450-5.

[86] Cade Metz, *Genius Makers: The Mavericks Who Brought AI to Google, Facebook, and the World* (Dutton, 2021).

[87] Tom Simonite, "Google's AI Guru Wants Computers to Think More like Brains," *Wired*, December 12, 2018, https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/.

[88] Nick Wallace, "EU's Right to Explanation: A Harmful Restriction on Artificial Intelligence," TechZone, January 25, 2017, http://www.techzone360.com/topics/techzone/articles/2017/01/25/429101-eus-right-explanation-harmful-restriction-artificial-intelligence.htm#.

[89] Mueller et al., "Explanation in Human-AI Systems."

[90] Bunn, "Working in Contexts for Which Transparency Is Important," 143.