

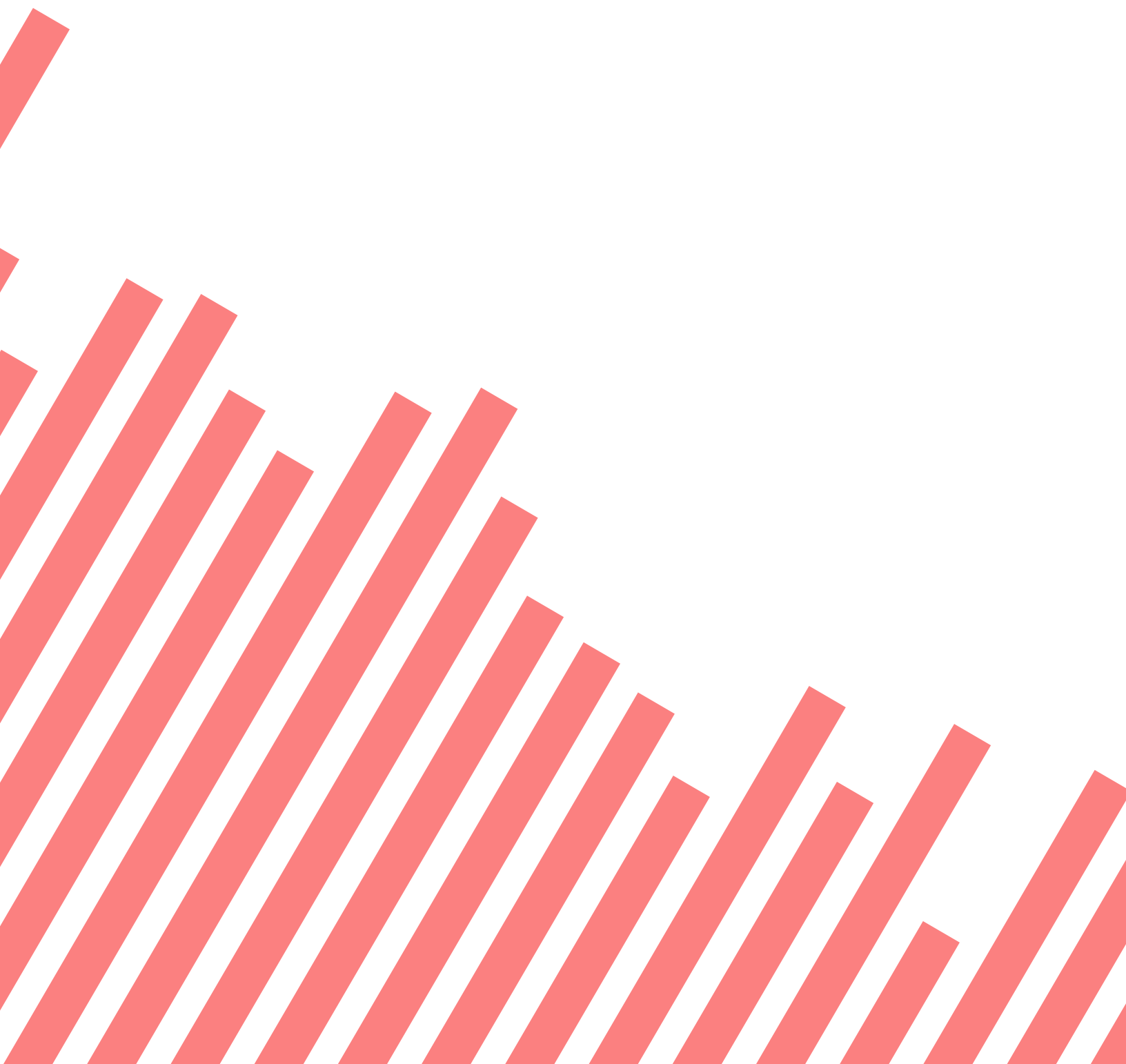
# Research Library Issues

RLI 299:

Ethics of Artificial Intelligence

2019

ASSOCIATION  
OF RESEARCH  
LIBRARIES



## In This Issue

<b>What Do Artificial Intelligence (AI) and Ethics of AI Mean in the Context of Research Libraries?</b>	<b>3</b>
The Context	
AI Ethics Initiatives in 2019	
AI Ethics and Research Libraries	
<b>Technology Innovation and AI Ethics</b>	<b>14</b>
Introduction	
Why AI Ethics?	
Technology Innovation and Wealth Distribution	
Data, Ethics, and Technocracy	
Future Humans	
What Should AI Ethics Look Like?	
<b>Explainable Artificial Intelligence</b>	<b>28</b>
Introduction	
Defining Explainable AI (XAI)	
European Union General Data Protection Regulation	
Opacity and Trust	
XAI Strategies, Techniques, and Processes	
XAI and Research Libraries	
AI-Authorship: An Explainability Sandbox	
Conclusion	
<b>Research Librarians as Guides and Navigators for AI Policies at Universities</b>	<b>47</b>
Introduction	
AI and Research Universities in the National and Global Context	
AI and Research Librarians	
A Role for Research Librarians in AI in University Research	
A Role for Research Librarians in AI in University Education	
A Role for Research Librarians in the Use of AI in University Administrative Systems	
Summary	
Further Reading	

## What Do Artificial Intelligence (AI) and Ethics of AI Mean in the Context of Research Libraries?

Mary Lee Kennedy, Executive Director, Association of Research Libraries

### The Context

The Association of Research Libraries (ARL) seeks to understand and engage the research library community and others in the research and learning ecosystem on the ethical implications of AI in the context of knowledge production, dissemination, and preservation. Furthermore, it seeks to inform the adoption of AI in research library operations, to help shape the research library workforce, and to advise and, as appropriate, help catalyze the services and programs that research libraries offer. With so much underway in the field of AI, there is a need for research libraries to act, starting with clarifying AI ethics policies, principles, and practices. This issue of *Research Library Issues (RLI)* opens up a conversation that ARL will continue to focus on in partnerships, and in formal and informal forums, particularly in the context of advocacy and public policy, institutional policies, research and learning community practices, and leadership development.

For the purpose of this issue of *RLI*, artificial intelligence is “the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.”<sup>1</sup> “Artificial intelligence” is not a new term. The first use of the term is attributed to John McCarthy at the 1956 Dartmouth Conference.<sup>2</sup> The concept of machine thinking is often attributed to Vannevar Bush’s seminal work in 1945, “As We May Think,” summarized so well by the editor as a paper that “**calls for a new relationship between thinking man and the sum of our knowledge**” [emphasis added].<sup>3</sup> And, Alan Turing is well known for his work during the Second World War on Enigma and the Bombe machine in laying the groundwork for machine learning.

The term “artificial intelligence” and its three primary related concepts (neural networks, machine learning, and deep learning) are used to varying degrees in the literature today. The presence of the term “artificial intelligence” in Google Books shows a distinct spike in the late 1980s and early 1990s, with increasing use of the term “machine learning” more recently, and, mildly, “deep learning” (see Figure 1). As access to large data sets grew, the potential for artificial intelligence, and therefore funding, also grew. Interestingly, there is a noticeable increase in the term “misinformation” during the rise in the mention of “machine learning.” It isn’t possible from this view alone to determine why there is a precipitous drop in the use of the term “artificial intelligence” in the 1990s, although one could hypothesize that, by then, we were all more aware of the different threads in artificial intelligence, and so used the distinctions like “machine learning” more often. Alternatively, it could be that with the growing interest in the field, publication moved from books to more timely forms of information sharing, such as journal articles. Though causality cannot be proven in any way through the Ngram Viewer, the rise in use of the word “misinformation” is not surprising, and reinforces the significant opportunity for and responsibility of our field.

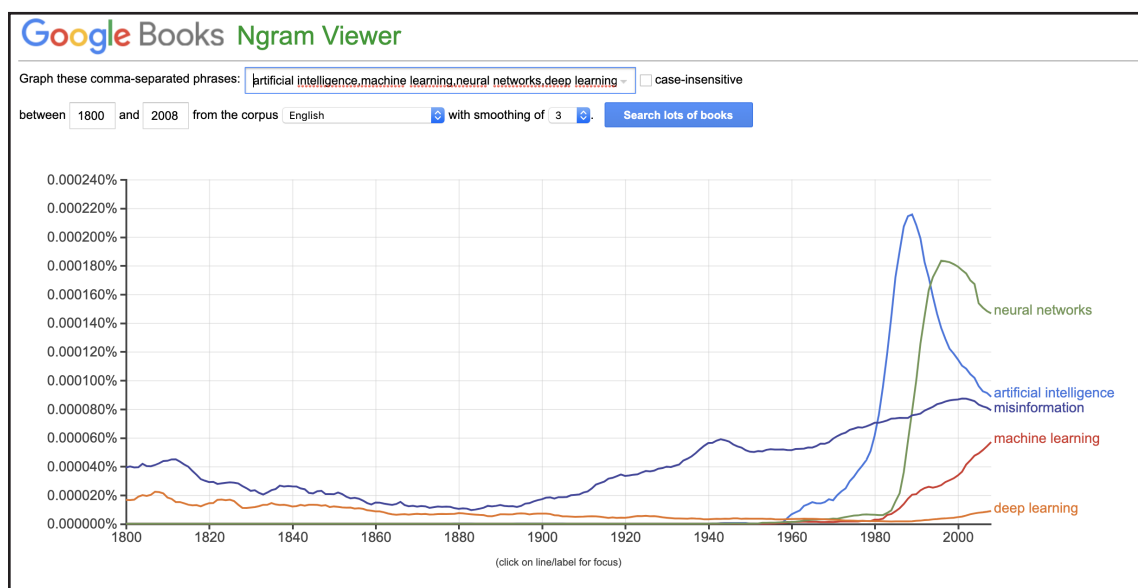


Figure 1



this discussion. A study by ARL, CNI, and EDUCAUSE over the next 18 months seeks to shed light on the critical technologies shaping research and learning (including AI), and the implications for research libraries as collaborative partners in the research enterprise.<sup>9</sup>

## **AI Ethics Initiatives in 2019**

There are significant AI ethics initiatives underway in 2019 both globally and nationally. This is a critical and opportune time for research libraries to assess and actively engage in informing the principles and practices of AI institutionally, in public policy, and in the research and learning community.

Following an initial draft in December 2018, in April 2019 the European Commission's High-Level Expert Group on Artificial Intelligence established seven essentials for achieving trustworthy artificial intelligence. The essentials are: human agency and oversight; robustness and safeness; privacy and data governance; transparency; diversity, non-discrimination, and fairness; societal and environmental well-being; and accountability.<sup>10</sup> On May 22, 2019, the Organisation for Economic Co-operation and Development (OECD) issued the "Recommendation of the Council on Artificial Intelligence,"<sup>11</sup> adopted by 42 countries, including Canada and the United States. On June 9, 2019, the G20 agreed on guiding principles for adopting artificial intelligence.<sup>12</sup> Although the principles among these bodies are not identical, they are more similar than different. Recently, France, Germany, and Japan agreed to jointly fund AI research that respects privacy and transparency.<sup>13</sup>

Closer to our members from Canada and the United States, on May 14, 2019, the Canadian Minister of Innovation, Science and Economic Development announced the Advisory Council on Artificial Intelligence. The purpose is to "advise the Government of Canada on building Canada's strengths and global leadership in AI, identifying opportunities to create economic growth that benefits all Canadians,

and ensuring that AI advancements reflect Canadian values.”<sup>14</sup> This followed a statement with France.<sup>15</sup> Further, Canada developed AI superclusters, including higher education institutions, to promote the development and use of AI.<sup>16</sup> The US National Institute for Standards and Technology (NIST) issued a request for information in May 2019 following the February 2019 United States Executive Order on “Maintaining American Leadership in Artificial Intelligence.” NIST’s charge is to “create a plan for federal engagement in the development of these standards and tools in support of reliable, robust and trustworthy systems that use AI technologies.”<sup>17</sup> In June 2019, the Office of the President issued *The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update*, including a section on ethics.<sup>18</sup> All of the above reinforces the undeclared but understood race for AI leadership in today’s world.

Declarations and recommendations are not limited to government bodies. The report of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems on ethically aligned design bases its recommendations on the principles of human rights, well-being, data agency, effectiveness, transparency, accountability, awareness of misuse, and competence. In their words:

Whether our ethical practices are Western (e.g., Aristotelian, Kantian), Eastern (e.g., Shinto, 墨家/School of Mo, Confucian), African (e.g., Ubuntu), or from another tradition, honoring holistic definitions of societal prosperity is essential versus pursuing one-dimensional goals of increased productivity or gross domestic product (GDP). Autonomous and intelligent systems should prioritize and have as their goal the explicit honoring of our inalienable fundamental rights and dignity as well as the increase of human flourishing and environmental sustainability.<sup>19</sup>

The Montreal Declaration for the responsible development of AI is being implemented with input from the general public.<sup>20</sup> Harvard University and MIT have a joint initiative to provide evidence-based

research to decision-makers in the private and public sectors, in order to advance the use of AI for public good.<sup>21</sup> Several institutions of higher education offer AI ethics courses to undergraduates and graduate students; examples include Stanford University,<sup>22</sup> Vanderbilt University,<sup>23</sup> University of Arizona,<sup>24</sup> as well as courses through edX, Coursera, and Udacity. Computer science program accreditation by ABET requires understanding of professional, ethical, legal, security, and social issues and responsibilities—including as ethics relate to AI.<sup>25</sup> In fact, an “AI ethics” online course search in Google returns over 35,000 results.

### **AI Ethics and Research Libraries**

In this global and national context of AI investments and adoption, this issue of *RLI* focuses on the relation of AI ethics and the role and potential roles of research libraries. A limited sampling highlights the broadening adoption of AI in research libraries. The University of Oklahoma<sup>26</sup> and other examples are highlighted in a 2019 issue of *Library Technology Reports* edited by Jason Griffey that opens with this compelling statement: “This issue of *Library Technology Reports* argues that the near future of library work will be enormously impacted and perhaps forever changed as a result of artificial intelligence (AI) and machine learning systems becoming commonplace.”<sup>27</sup>

Now is the time for research libraries to collectively understand and address a host of ethical questions for research institutions, public policy, and more specifically for research library leaders in institutional and public policy, so that research libraries will continue to serve as trusted advisors to our users, and as responsible collectors, disseminators, and preservers of knowledge. To help frame our thinking, we invited three individuals to share their expertise and recommendations with us.

**Sylvester Johnson**, the founding director of the Center for Humanities and the assistant vice provost for the humanities at

Virginia Tech, focuses on the role of ethics in innovation in the first article in this issue. AI, like other influential technologies, can be a force for innovation, and is known to have harmful as well as helpful implications. Johnson highlights the undeniable moment in which technologies are raising fundamental ethical questions about humanity, including how we want to inhabit the world that we are creating. With information at the core, he lays out opportunities and challenges for research libraries.

Within the broad context of policy and principles, there is an opportunity for research libraries to make a difference today—explainable artificial intelligence (XAI). In the second article, **Michael Ridley**, Librarian Emeritus at University of Guelph, PhD candidate at Western University, and postgraduate affiliate at Vector Institute, defines XAI, and then situates it in the context of privacy, opacity, and trust. He advances our understanding of XAI by outlining strategies, techniques, and processes. He concludes by squarely putting the opportunity on research libraries “to shape the development, deployment, and use of intelligent systems in a manner consistent with the values of scholarship and librarianship” with XAI as one of the most important ways to do so.

**Geneva Henry**, dean of Libraries and Academic Innovation at The George Washington University, ties it all together for us with an article on the role of the research library in formulating and implementing institutional policy based on the needs of the users, and in the context of public policy. Starting out with an assessment of national investments in AI, Henry emphasizes the role of policies that promote ethically responsible practice. Her article outlines ways in which research libraries are answering and could answer the key question posed by Brundage and Bryson: it “is not whether AI will be governed, but how it is currently being governed, and how that governance might become more informed, integrated, effective, and anticipatory.”<sup>28</sup>

I hope you will discover new knowledge and urgency in the articles published here. Please contact me or any of the authors with questions or suggestions.

## Endnotes

1. *Lexico*, s.v. “artificial intelligence,” accessed September 12, 2019, [https://www.lexico.com/en/definition/artificial\\_intelligence](https://www.lexico.com/en/definition/artificial_intelligence).
2. James Moor, “The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years,” *AI Magazine* 27, no. 4 (Winter 2006): 87–91, <https://pdfs.semanticscholar.org/d486/9863b5da0fa4ff5707fa972c6e1dc92474f6.pdf>.
3. Vannevar Bush, “As We May Think,” *The Atlantic*, July 1945, <https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/>.
4. Klaus Schwab, “The Fourth Industrial Revolution: What It Means, How to Respond,” World Economic Forum, January 14, 2016, <https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/>.
5. “Artificial Intelligence and Robotics,” interactive data visualization, World Economic Forum and University of Cambridge, accessed September 12, 2019, <https://intelligence.weforum.org/topics/a1Gb0000000pTDREA2?tab=publications>.
6. Vesselina Stefanova Ratcheva and Till Leopold, “5 Things to Know about the Future of Jobs,” World Economic Forum, September 17, 2018, <https://www.weforum.org/agenda/2018/09/future-of-jobs-2018-things-to-know/>.
7. “5 Ways Artificial Intelligence May Influence Higher Education Admissions & Retention,” Wiley Education Services, accessed September 12, 2019, <https://edservices.wiley.com/artificial-intelligence-in-higher-ed/>.

8. Justin Klutka, Nathan Ackerly, and Andrew J. Magda, *Artificial Intelligence in Higher Education: Current Uses and Future Applications* (Louisville, KY: Learning House, November 2018), <https://www.learninghouse.com/knowledge-center/research-reports/artificial-intelligence-in-higher-education/>.
9. Mary Lee Kennedy, “ARL, CNI, EDUCAUSE Form Strategic Partnership to Advance Research Libraries’ Impact in a World Shaped by New Technologies,” Association of Research Libraries, August 16, 2019, <https://www.arl.org/news/arl-cni-educause-form-strategic-partnership-to-advance-research-libraries-impact-in-a-world-shaped-by-new-technologies/>.
10. “Ethics Guidelines for Trustworthy AI,” European Commission, April 8, 2019, <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.
11. “Recommendation of the Council on Artificial Intelligence,” Organisation for Economic Co-operation and Development, adopted May 21, 2019, <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.
12. Masumi Koizumi, “G20 Ministers Agree on Guiding Principles for Using Artificial Intelligence,” *Japan Times*, June 8, 2019, <https://www.japantimes.co.jp/news/2019/06/08/business/g20-ministers-kick-talks-trade-digital-economy-ibaraki-prefecture/>.
13. David Matthews, “New Research Alliance Cements Global Split on AI Ethics,” *Times Higher Education (THE)*, August 15, 2019, <https://www.timeshighereducation.com/news/new-research-alliance-cements-global-split-ai-ethics>.
14. “Advisory Council on Artificial Intelligence,” Government of Canada, modified May 14, 2019, <http://www.ic.gc.ca/eic/site/132.nsf/eng/home>.
15. “Canada-France Statement on Artificial Intelligence,” Government of Canada, modified June 7, 2018, <https://www.international.gc.ca/>

[world-monde/international\\_relations-relations\\_internationales/europe/2018-06-07-france\\_ai-ia\\_france.aspx?lang=eng](http://world-monde/international_relations-relations_internationales/europe/2018-06-07-france_ai-ia_france.aspx?lang=eng).

16. “Government of Canada Creates Advisory Council on Artificial Intelligence,” Government of Canada, May 14, 2019, <https://www.canada.ca/en/innovation-science-economic-development/news/2019/05/government-of-canada-creates-advisory-council-on-artificial-intelligence.html>.
17. “NIST Requests Information on Artificial Intelligence Technical Standards and Tools,” National Institute for Standards and Technology, May 1, 2019, <https://www.nist.gov/news-events/news/2019/05/nist-requests-information-artificial-intelligence-technical-standards-and>.
18. Select Committee on Artificial Intelligence of the National Science & Technology Council, *The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update* (Washington, DC: Executive Office of the President of the United States, June 2019), <https://www.nitrd.gov/news/National-AI-RD-Strategy-2019.aspx>.
19. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, IEEE, 2019, <https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf>.
20. “Artificial Intelligence,” Université de Montréal, accessed September 12, 2019, <https://www.umontreal.ca/en/artificialintelligence/>.
21. “Ethics and Governance of Artificial Intelligence,” Berkman Klein Center for Internet & Society at Harvard University and MIT Media Lab, accessed September 12, 2019, <https://www.media.mit.edu/groups/ethics-and-governance/overview/>.
22. “CS122: Artificial Intelligence—Philosophy, Ethics, and Impact,” Stanford University, Fall 2014, <https://web.stanford.edu/class/cs122/>.

23. “The Ethics of Artificial Intelligence: A University Course,” Vanderbilt University, Spring 2019, <https://my.vanderbilt.edu/aiethics/>.
24. “Undergraduate Philosophy Courses,” University of Arizona, accessed September 12, 2019, <https://philosophy.arizona.edu/undergraduate-courses>.
25. “Criteria for Accrediting Computing Programs, 2018–2019,” ABET, accessed September 12, 2019, <https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-computing-programs-2018-2019/>.
26. Jeffrey R. Young, “Bots in the Library? Colleges Try AI to Help Researchers (But with Caution),” *EdSurge*, June 14, 2019, <https://www.edsurge.com/news/2019-06-14-bots-in-the-library-colleges-try-ai-to-help-researchers-but-with-caution>.
27. Jason Griffey, ed., “Artificial Intelligence and Machine Learning in Libraries,” *Library Technology Reports* 55, no. 1 (January 2019), <https://doi.org/10.5860/ltr.55n1>.
28. Miles Brundage and Joanna Bryson, “Smart Policies for Artificial Intelligence,” preprint, submitted August 29, 2016, 12, <https://arxiv.org/abs/1608.08196>.

© 2019 Mary Lee Kennedy



This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>.

**To cite this article:** Mary Lee Kennedy. “What Do Artificial Intelligence (AI) and Ethics of AI Mean in the Context of Research Libraries?” *Research Library Issues*, no. 299 (2019): 3–13. <https://doi.org/10.29242/rli.299.1>.

## Technology Innovation and AI Ethics

**Sylvester A. Johnson**, Assistant Vice Provost for the Humanities and Director of the Center for Humanities, Virginia Tech

### Introduction

In 2017, Saudi Arabia granted citizenship to a machine, a humanoid robot powered by artificial intelligence (AI) and named Sophia. This woman-gendered robot is manufactured by the Hong Kong-based Hanson Robotics Corporation and it is machine-learning technology that enables her to deliver scripted speech and to participate in spontaneous conversation with humans, complete with facial gestures, intonation, and other forms of body language. Sophia had just delivered a speech at the nation's Future Investment Initiative summit, to which Saudi Arabia had invited hundreds of global investors to consider leveraging the financial growth opportunities the nation is charting for its future. Following Sophia's speech, it was announced that the government had granted her citizenship. Sophia responded with delight, even pondering the possibility of voting and attending college one day.

It seems undeniable that technology innovation is broaching fundamental questions about humanity and ethics. In the wake of Sophia's citizenship announcement, a mix of fascination and dissent emerged. Many people were quite amused that this humanoid AI robot could be so charming. Others lamented the fact that Saudi Arabia had granted citizenship rights to a machine while denying the same to millions of human immigrants. Still others noted that Sophia addressed an audience of elite men while unveiled, whereas human women in Saudi Arabia are traditionally required to veil in public. Amidst the various responses, one thing was certain: granting citizenship to an intelligent machine was a sure sign that AI technology is as much a social issue as it is a technical one.

In a prescient commencement address at Northwestern University back in 2015, IBM's CEO Ginny Rometty identified an emerging paradigm shift, declaring that "the dawn of a new era" is upon us, one in which "every important decision" of humankind will be made not by humans alone, but by human-machine alliances powered by "cognitive computing" systems to enable outcomes beyond anything humans might accomplish on their own.<sup>1</sup> Rometty was right to recognize that such themes as creativity, research, and culture have been traditionally conceived as uniquely human accomplishments in the past and are increasingly being performed by machines and humans working in concert. For several years now, IBM's Watson AI system has been working with human oncologists at the Memorial Sloan Kettering Cancer Center to learn how to develop treatments for cancer. Watson is also being used to assist decision-making in other domains, such as finance, marketing, and concierge services.

“Granting citizenship to an intelligent machine was a sure sign that AI technology is as much a social issue as it is a technical one.”

Although Hanson Robotics's Sophia and IBM's Watson can be competent at very specific tasks such as having a friendly conversation (Sophia) or reading and understanding thousands of articles on a given subject (Watson), humans still reign at so-called general intelligence. We think nothing of the fact that a single human might be equally adept at cooking, composing music, reading a data chart, and building furniture. This is just common sense. This range of ability simply does not exist with artificial intelligence. At least not yet. But in July of 2019, the Microsoft Corporation formally partnered with the formerly nonprofit OpenAI to inaugurate a new collaboration that aims to build the world's first machine intelligence capable of human-level general intelligent comprehension and skill.<sup>2</sup> If successful, such an AI system would be as adept at playing chess and creating recipes as it would be at curing cancer, analyzing foreign policy, planning urban development, and deriving practical solutions to address climate change. This would be a true know-it-all, capable of learning anything on its own without

having to be programmed. The scale of this machine intelligence would far outpace the capacity of any biological human, just as a simple calculator can outperform any human at solving math problems.

Technology innovation is creating immense opportunities to improve the lives of people throughout the world. As is especially evident through advances in artificial intelligence, this innovation is also producing startling quandaries that at one time seemed far-fetched and fictitious, but that now raise ethical challenges for the present and future of humanity.

### **Why AI Ethics?**

As daunting as the technical questions are for fulfilling the vision of an AI-driven world, it appears that the ethics of governing innovation will be even harder. How should we manage technology—how do we shape outcomes, processes, and consequences—to ensure that human society is not only sustainable but also thriving? More bluntly, how will we create a future we actually want to inhabit, rather than one defined by destitution, technological cataclysm, and inhumane conditions? The answers to such questions are not simply technical; rather, they are profoundly humanistic and comprehensive. The judgments and decisions that will shape our human future are ultimately ethical in nature. They mandate consideration of social benefits and costs, of material advantage and disadvantage, and of security, wealth, and well-being.

In an astute article about the future of a digital society, Palmer Group CEO Shelly Palmer voiced a similar concern, explaining that the choice we face is not merely about opting in or opting out of “privacy,” but “about our economic sovereignty and our national security.”<sup>3</sup> There is abundant concern that resonates with such cautionary voices. Many readers will recall that in 2016, several entrepreneurs and scientists from Elon Musk to Stephen Hawking signed an open letter urging governments to ban weaponized forms of AI until experts

have developed a reliable way to control such technology. In 2018, Google responded to protests from their own employees by agreeing to refrain from developing AI for weapons systems and other forms of destruction.

## **Technology Innovation and Wealth Distribution**

There can be little question that technology innovation is driving transformative changes nationally and globally. Joseph Stiglitz, former chief economist for the World Bank, observes that the economic growth resulting from technology will create unprecedented wealth in the years to come, albeit through drastically uneven distribution and with social implications that will require judicious foresight and humanistic guidance.<sup>4</sup> Consider that in just the next 10 years, digital technology is estimated to add around \$100 trillion (net) in GDP to the global economy. Of that amount, AI alone will be responsible for about \$15 trillion.<sup>5</sup> It is a staggering figure, but that's just the next decade. What about the next quarter-century? No less a leader than Kai-Fu Lee, the CEO of Sinovation Ventures and formerly president of Google China, has argued that AI will produce a scale of inequality that will create a gaping wealth divide between regions of the world as well as within individual nations. Without some drastic intervention, this inequality will escalate at a speed that previous analysts have scarcely imagined.<sup>6</sup> This is because the massive wealth that AI generates is concentrated into the hands of an increasingly smaller portion of humanity. This is happening at a time of increasing precarity for the middle-class—inflation-adjusted, real wages are stagnant or declining—and dissipating political support for the poor.

Were living-wage jobs to decrease rapidly due to AI automation (consider that AI is already replacing humans in finance and healthcare), millions more people of a previously middle- or upper-class would be plunged into underemployment, unemployment, and poverty. Stiglitz has urged that the current relationship between capital and politics, moreover, has already created an environment

that cultivates more scorn and contempt than compassion for the poor, among whom racial minorities are disproportionately represented. In the United States, particularly, federal and state policies have directed billions of dollars more toward prisons and militarized policing of the poor than toward education and healthcare for those same people. As a result, the country commands the number one spot as the world's top incarceration nation. Absent a drastic shift in American politics, it is difficult to imagine how a technologically driven, rapid increase in AI labor-automation would not end disastrously for most people. The rise of nationalist political parties on a global scale, moreover, that target the poor, immigrants, and racial minorities as an existential threat does not bode well for a future in which accelerating inequality will demand transnational synergies and collaboration to ensure a viable existence for humankind.

### **Data, Ethics, and Technocracy**

Technology innovation, of which AI is an especially powerful example, has proceeded most vigorously through information science. This might be more familiar to contemporary readers as “data science.” This latter term has become both a mantra and a chief paradigm for business, culture, entertainment, and security. It was only a few decades ago that most companies had never heard of a chief information officer (CIO). Today, executive administration of an organization's information is as standard as financial accounting. Information, in fact, is now the most valuable asset a company possesses.

Data science uses software (algorithms) to interpret massive data sets (the equivalent of millions of DVDs-worth of information, for instance) to produce insights into the real world. Such large sets of data are beyond what any human could possibly handle. But algorithms quickly churn through thousands of data points on a single individual to discern patterns of behavior so well that the software can reliably predict what people will do. Will you want to purchase a new hat

or lawn mower next month? What type of car will you buy in two years? Technology companies probably know before you do and can sell advertising to the highest bidder standing by to translate your purchasing power into their future revenue.

The world learned from the 2016 US presidential election and from England's Brexit campaign that data science in the form of psychographics might also be used to modify people's behavior. This was the basis for Cambridge Analytica's business model. Using 5,000 data points on a given individual, the company had developed over many years effective means of directing the decisions that targeted voters would make at the polls.<sup>7</sup> Entities ranging from governments to schools to private corporations to law enforcement enjoy unprecedented access to previously unimaginable volumes of data about people throughout the world today. This fact alone poses an immense ethical issue: how much data is too much for any entity to possess? Should data sets be classified as public to prevent them from being monetized? If monetizing data represents a viable path for ethical outcomes in humanity's future, should individuals benefit financially from the use of their data? These difficult questions defy easy answers; but they must be met with deliberation at both practical and regulatory levels if we are to avoid the most undesirable consequences.

As it happens, information is also the central concern of library science and of the educational domain more broadly. This poses both a special opportunity and a perplexing challenge for academic libraries specifically and for the world's educational institutions more generally. The opportunity rests largely with the fact that libraries occupy the center of gravity in a technocratic society because they manage the most valuable asset category the world has known—data, “the oil of the digital era.”<sup>8</sup> This also implies that libraries are uniquely positioned to leverage innovation for enhancing the delivery of information to a broad population of learners. The challenge, by contrast, is perhaps best demonstrated by the monetization of digital information.

This is precisely the nut to be cracked in the quandary over the Elsevier corporation, which is self-designated as an “information analytics” (more commonly termed data analytics) company. At a practical level, Elsevier controls access to most of the world’s published research. Like other data analytics companies, Elsevier is able to leverage and monetize the insights gained from mining massive amounts of data about users. By an accident of history, academic libraries find themselves obligated to expend billions of dollars for contracts with Elsevier to ensure that the consumers of information (students, faculty, researchers, etc.) can access knowledge in the form of academic publishing. As libraries run up against the limits of their financial resources, they will have to consider what role they will play in the information economy.

Librarians have begun grappling with the ethical nature of this situation and with the imperative of structuring a viable and sustainable future for delivering information.

A major aspect of this ethical challenge that technology innovation is raising for libraries can be put more sharply: what relationship should exist between information (in the form of scholarly research, for instance) and markets? And never mind the adage that “information should be free.” In the real world of employees, book purchases, journal subscriptions, capital assets, and institutional finance, such a refrain merely dodges the question. Will academic libraries remain conduits for the behavioral data their users are generating? Should libraries also participate in deriving insights from user data and monetize those insights for ethical ends? Do there exist inherent tensions in this enterprise?

As daunting as these challenges seem, the trajectory of technology innovation appears set to deliver even more complicated quandaries. As machine intelligence achieves greater capacity to read and understand expert material, at what point will AIs be recruited to write scholarly papers on subjects ranging from history to psychology to

economics to oncology and computer science? IBM's Project Debater is already capable of ingesting thousands of articles about a given subject, understanding the content, and debating a human by discerning that person's argument to generate a counter-argument rooted in information-rich analysis.<sup>9</sup> It is already the case that AIs can read in a few hours more than any single human could possibly read in their entire lifetime. So, there is certainly a compelling argument to be made that scholarship produced by AIs has a role to play in advancing expert knowledge to promote understanding, analysis, and innovation.

“Now is the time for a broad array of experts to anticipate new directions in technology innovation in order to begin shaping an ethical and sustainable future rooted in equitable outcomes.”

What would authorship mean in such a scenario? Should an AI be legally recognized as an author, particularly when no human could possibly generate the robust analysis and writing such a system might create? If AIs create publications and disseminate knowledge that relies on the research produced by academics (this reliance on expertise disseminated through academic publishing is the current model), then who should benefit from

any monetization of such authorship? Machine-learning systems are already leveraging existing research (generated by humans) to derive insights for treating or curing disease. Who will own the capital (servers, algorithmic design, cloud-based services, data sets, and so forth) that is the basis for the digital domains of technology innovation as AIs join the rank of academic publishers?

It might be tempting to simply hope such developments never occur. And yet, Microsoft and OpenAI have already forged an alliance to make this scenario look like child's play. As existing AIs can already write poetry, short stories, and newspaper articles, it is certain that even a minimally successful product that barely approximates the goal of artificial general intelligence (AGI) could mean the irreversible transformation of the expert knowledge economy. It should be clear

that hoping against the future is not a strategy. Instead, now is the time for a broad array of experts to anticipate new directions in technology innovation in order to begin shaping an ethical and sustainable future rooted in equitable outcomes.

## Future Humans

Perhaps the horizon of technology innovation will increasingly be shaped by developments in human enhancement or human engineering. Advances in human-machine combining (cybernetics) and genetic engineering promise to create radical changes to human society and unprecedented questions of ethics, equity, and accountability that will easily match or exceed those being generated by AI. Today, every major military industrial state is racing to develop capacities in military soldiers that surpass those of unmodified humans. These efforts include exoskeletons, drug enhancements, brain implants, and “smart” (AI-driven) prostheses; such efforts would permit soldiers to carry heavier loads greater distances, control tools or weapons by thinking,

“As our global society increasingly recognizes that technology is not merely technical but also societal and human-oriented, new doors of opportunity are opening for humanists to take leadership of the most important efforts that might shape the future of society.”

process information more quickly than a normal human by interfacing with an AI system, or function on alert for days without sleeping. Given the high stakes of military dominance for which the world’s most powerful military nation-states

compete, there is every reason, as well, to expect genetic engineering to emerge in military applications on a global scale.

Medical therapies constitute arguably the most compelling motivation for aggressively pursuing human enhancement. It is one thing, after all, to rationalize modifying humans for warfare, which inherently involves killing and destruction. It is quite another to justify modifying humans

as a means of preserving life and restoring capacities; these efforts can enable patients to regain speech and motor function, for instance. Even this medical context for enhancement is generating well warranted concerns about ableism and eugenics, particularly as the meaning of a “normal body” or “normal capacity” is reshaped by this technology. It seems unlikely on ethical grounds, however, that such valid concerns will be used to deny everyone even the possibility of regaining the ability to walk again or to have impaired vision or hearing restored through technological enhancements as the development of these technologies continues to advance.

All of this means that technology innovation is on pace to reshape the future of humanity in deeply consequential ways, including at the foundational level of what it means to be (a) human.

### **What Should AI Ethics Look Like?**

As our global society increasingly recognizes that technology is not merely technical but also societal and human-oriented, new doors of opportunity are opening for humanists to take leadership of the most important efforts that might shape the future of society. The University of Oxford announced with great fanfare in June of 2019 that Blackstone CEO Stephen Schwarzman had gifted more than \$188 million to fund a humanities center housing a new institute for AI ethics. The billionaire-philanthropist had previously donated \$350 million to MIT to create an institute-wide “College of AI” that will emphasize the role of the liberal arts and human sciences. In 2019, Stanford University launched a new institute harnessing university-wide efforts to support human-centered AI, placing at the helm a philosopher and a computer scientist.

We are witnessing growing efforts to ensure that technology serves human interests through regulatory efforts, ethical frameworks, and more comprehensive education. “Public interest technology” is among the key growth areas devoted to ensuring that social justice and

equity guide the development and implementation of technology. Since 2016, philanthropic foundations have devoted millions of dollars to support this new approach for American universities to prepare a new generation of “technologists” to work in civil service, education, and a full range of humanistic endeavors. The daunting challenges of AI have motivated major technology corporations such as Apple, Google, and Microsoft to emphasize fairness, ethics, and humanistic approaches to innovation. The London-based company DeepMind has made AI ethics central to the guidance of its technology. Since 2018, Google has published AI principles underscoring their commitment to fairness and avoiding developing AI weapons and other harmful forms of AI technology. In 2019, China’s Beijing Academy of Artificial Intelligence published ethical guidelines emphasizing fairness and sustainability. The World Economic Forum has likewise articulated ethical guidelines for the use of technology.

“After decades of worries that the popularity of science and technology paradigms threaten humanistic learning and scholarship, it is now becoming evident that unique opportunities are emerging to demonstrate why humanistic expertise and informed considerations of the human condition are essential to the very future of humanity in a technological age.”

Notwithstanding this important beginning, bringing ethical governance to technology will require a thoroughgoing transformation of humanities leadership. Colleges and universities will need to invest greater resources in humanistic programs of study. Humanistic disciplines must focus more urgently on recruiting and producing far more racial minorities and women in technology. Because technology innovation will bring massive changes to our democratic institutions and social systems, future technologists will have to include people with expertise in the human condition, policy, and social services.

Equally important will be transdisciplinary communities of research and collaboration that must provide teams of diverse talent and expertise to guide the use of AI in higher education, law enforcement,

medicine, finance, and warfare. As things currently stand, there exists no regulatory framework for governing technology innovation. The good news is that the challenges posed by the ethical guidance of AI and other forms of technology innovation will require our social institutions to embrace new forms of leadership from humanities experts. After decades of worries that the popularity of science and technology paradigms threaten humanistic learning and scholarship, it is now becoming evident that unique opportunities are emerging to demonstrate why humanistic expertise and informed considerations of the human condition are essential to the very future of humanity in a technological age.

## Endnotes

1. Virginia Rometty, “Northwestern University Commencement Address,” C-SPAN, June 19, 2015, video, 14:33, <https://www.c-span.org/video/?326217-1/virginia-rometty-commencement-address-northwestern-university>.
2. Kelsey Piper, “Microsoft Wants to Build Artificial General Intelligence: An AI Better than Humans at Everything,” *Vox*, July 22, 2019, <https://www.vox.com/2019/7/22/20704184/microsoft-open-ai-billion-investment-artificial-intelligence>.
3. Shelly Palmer, “Governing a Digital Democracy: Unanswered Questions,” LinkedIn, February 10, 2019, <https://www.linkedin.com/pulse/governing-digital-democracy-unanswered-questions-shelly-palmer/>.
4. Joseph E. Stiglitz, *People, Power, and Profits: Progressive Capitalism for an Age of Discontent* (New York: W. W. Norton, 2019), 35–38, 46.
5. Barry B. Hughes et al., “ICT/Cyber Benefits and Costs: Reconciling Competing Perspectives on the Current and Future Balance,” *Technological Forecasting and Social Change* 115 (February 2017): 117–130, <https://doi.org/10.1016/j.techfore.2016.09.027>; Frank Holmes, “AI Will Add \$15 Trillion to the World Economy by 2030,” *Forbes*, February 25, 2019, <https://www.forbes.com/sites/greatspeculations/2019/02/25/ai-will-add-15-trillion-to-the-world-economy-by-2030/>. PwC Global

puts the figure in the same range, at \$15.7 trillion, exceeding the current combined GDP of China and India. See *Sizing the Prize: What's the Real Value of AI for Your Business and How Can You Capitalise?*, PwC Global, 2017, <https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf>. The McKinsey Global Institute estimates a slightly lower but similar figure, \$13 trillion. See Jacques Bughin et al., *Notes from the AI Frontier: Modeling the Impact of AI on the World Economy*, McKinsey Global Institute, September 2018, <https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Notes%20from%20the%20frontier%20Modeling%20the%20impact%20of%20AI%20on%20the%20world%20economy/MGI-Notes-from-the-AI-frontier-Modeling-the-impact-of-AI-on-the-world-economy-September-2018.ashx>.

6. Kai-Fu Lee, *AI Superpowers: China, Silicon Valley, and the New World Order* (Boston: Houghton Mifflin Harcourt, 2018), 170–173.
7. Varoon Bhashyarkar, “Psychometric Profiling: Persuasion by Personality in Elections,” *Our Data Our Selves*, Tactical Tech, May 18, 2018, <https://ourdataourselves.tacticaltech.org/posts/psychometric-profiling/>; Terrell McSweeney, “Psychographics, Predictive Analytics, Artificial Intelligence, & Bots: Is the FTC Keeping Pace?,” *Georgetown Law Technology Review* 2, no. 2 (July 2018): 514–530, <https://georgetownlawtechreview.org/psychographics-predictive-analytics-artificial-intelligence-bots-is-the-ftc-keeping-pace/GLTR-07-2018/>.
8. “The World’s Most Valuable Resource Is No Longer Oil, but Data,” *The Economist*, May 6, 2017, 9, <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data>.
9. “Project Debater,” IBM, accessed September 10, 2019, <https://www.research.ibm.com/artificial-intelligence/project-debater/>.

© 2019 Sylvester A. Johnson



This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>.

**To cite this article:** Sylvester A. Johnson. “Technology Innovation and AI Ethics.” *Research Library Issues*, no. 299 (2019): 14–27. <https://doi.org/10.29242/rli.299.2>.

## **Explainable Artificial Intelligence**

**Michael Ridley**, Librarian Emeritus, University of Guelph; PhD Candidate, Western University; and Postgraduate Affiliate, Vector Institute

### **Introduction**

Algorithmic decision-making, enabled by machine learning, is ubiquitous, powerful, often opaque, sometimes invisible, and, most importantly, consequential. Machine learning is embedded in many information tools and systems, central to numerous research methods, and pervasive in the applications of everyday life. Safiya Noble emphasizes the critical nature of artificial intelligence (AI) by observing that it will become “a major human rights issue in the twenty-first century.”<sup>1</sup> As with nearly all aspects of contemporary life, AI is having a profound influence on research libraries, scholarly communication, and key functions of the academy.

Because “authority is increasingly expressed algorithmically,”<sup>2</sup> it is crucial that this authority be interrogated and assessed with the same rigor and appropriate methods relevant to all aspects of the academic mission. Machine learning and deep learning are potent technologies that will be utilized to great advantage. 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.”<sup>3</sup> Hence the critical importance of “explainable artificial intelligence” (XAI) and its two pillars: trust and accountability.

XAI is a diverse set of strategies, techniques, and processes that render AI systems interpretable and accountable. While some XAI approaches are highly technical, involving the perturbation of individual features in multi-layer neural network models, others are broad social and political policies enacted through regulation or legislation. Whatever the approach, XAI emphasizes explainability as an essential requirement for a technology that has for too long been defined by its opacity and what Frank Pasquale calls “remediable incomprehensibility.”<sup>4</sup>

The use and development of machine learning applications in research libraries will only continue to grow in volume and influence. As AI reconfigures much of scholarly communications, it will be essential that libraries, and their users, have trust in the cognitive delegation of many tasks and processes. Mariarosaria Taddeo notes that “delegation without supervision characterises the presence of trust.”<sup>5</sup> Approaching that state will require artificial intelligence to exhibit, and be open to, new levels of transparency and accountability. One critical element of that is explainability.

“As AI reconfigures much of scholarly communications, it will be essential that libraries, and their users, have trust in the cognitive delegation of many tasks and processes.”

### **Defining Explainable AI (XAI)**

The US Defense Advanced Research Projects Agency (DARPA) definition of XAI is widely referenced. The purpose of XAI is to enable human users “to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.”<sup>6</sup> To this user-centric XAI definition, DARPA adds the expectation that AI systems “will have the ability to explain their rationale, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future.”<sup>7</sup> Examining these definitions yields both insights and complications.

The user-centric definition has three key concepts: understanding, trust, and management. Understanding can mean a range of ideas from simple awareness or acceptance to acknowledgement and finally to detailed knowledge. While the idea of trust seems straightforward, the modifier “appropriately” suggests a conditional situation where the granting of trust is contextual. Managing indicates a relationship between the user and the AI and implies that the human user is, or should be, in a position of reasonable control. However, referencing AI systems as “partners” suggests a more cooperative and quasi-independent relationship.

The system-centric definition also has three key concepts: rationale, strengths and weaknesses, and future behavior. The rationale could pertain to the purpose of the AI, the logic of its model, a justification for its actions, or its application in specific situations. The disclosure of strengths and weaknesses indicates a level of openness and transparency that would make obvious system limitations and key assumptions. It also seems likely to conflict with trade secrecy, intellectual property issues, and data privacy. The emphasis on future behavior recognizes that AI will be an ongoing part of everyday life, hence the need for predictability and consistency. It also implies that AI will be subject to longitudinal evaluations to ensure levels of performance.

### **European Union General Data Protection Regulation**

It is difficult to overestimate the impact of the European Union (EU)'s 2018 General Data Protection Regulation (GDPR) on XAI. The GDPR's "right to explanation" regarding algorithmic decisions is having a global reach (the so-called "Brussels effect"), causing debate and regulatory review well beyond the EU. While interpretability has always been a concern in computer science, the GDPR has refocused this issue as an explainability problem and made it a public policy question.

The explanatory requirements in the GDPR are actually quite narrow, but their impact has been much broader, with jurisdictions as diverse as Canada and the City of New York developing impact assessment protocols with respect to algorithmic decisions that include explainability requirements. As seen with the "right to be forgotten," international legislation or regulation can have a profound effect on national affairs. The global nature of digital technologies is a reminder that monitoring the policy agendas of other jurisdictions is important.

## Opacity and Trust

Why do we need an explanation for how AI works? Geoffrey Hinton, often referred to as the godfather of deep learning and neural networks, observes, “A deep-learning system doesn’t have any explanatory power...the more powerful the deep-learning system becomes, the more opaque it can become.”<sup>8</sup> Despite this, Hinton has been critical of requirements that AI should explain itself and insists performance should be the key measure of trust. After all, humans can’t provide explanations for many of their actions or decisions, why expect AI to do otherwise?

While Hinton may discount the importance of, or even the need for, an explanation, psychologists and cognitive scientists do not. Explanations are “more than a human preoccupation—they are central to our sense of understanding, and the currency in which we exchange beliefs.”<sup>9</sup> There is an extensive literature on both the power and the failings of AI. Examples of discrimination and unfairness are matched by extraordinary advances and success. However, it is exactly for these reasons that the opacity, complexity, and consequential nature of AI drives the need for trust and elevates explanation as a key antidote.

“As academic libraries increasingly acquire and develop algorithmic decision-making systems and services in support of scholarly communications and the operation of the library, they must do so in a manner that insists on interpretability and explanation.”

What is a good or satisfactory explanation? For whom is the explanation provided, in what context, with what, if any, evidence, and presented in what manner? An explanation should be able to address “how” (inputs, output, process), “why” (justification, motivation), “what” (awareness that an algorithmic decision-making system exists), and the “objective” (design, maintenance).<sup>10</sup> In the context of opaque systems, an explanation should be:

- “(1) model-agnostic, so it can be applied to any black box model;
- (2) logic-based, so that explanations can be made comprehensible to humans with diverse expertise, and support their reasoning;
- (3) both local and global, so it can explain both individual cases and the overall logic of the black-box model;
- (4) high-fidelity, so it provides a reliable and accurate approximation of the black box behavior.”<sup>11</sup>

A more holistic view would include explanations that consider the data used for training and decision-making, the computational environment utilized, the context of the algorithmic design and deployment, and those responsible for its operation and use (that is, a sociotechnical analysis).<sup>12</sup> Technical explanations are required for those involved in system design and performance testing while accessible explanations are needed for those affected by algorithms. In the latter context, a good explanation is contrastive (“why P not Q?”), selective (only certain evidence is required not a complete explanation), and social (a dialogue, interactive, contextual).<sup>13</sup>

As academic libraries increasingly acquire and develop algorithmic decision-making systems and services in support of scholarly communications and the operation of the library, they must do so in a manner that insists on interpretability and explanation. To do anything less is an unknowing delegation to technology and an abrogation of scholarly rigor.

### **XAI Strategies, Techniques, and Processes**

Approaches to XAI can be broadly categorized as proofs, validations, and authorizations. Within these categories are numerous explanatory practices, which are contextual, system or model dependent, and audience specific.

## *Proofs*

Proofs as explanations are testable, demonstrable, traceable, and unambiguous. In the context of AI, they pertain primarily to rule-based expert systems or systems that use decision trees (that is, an explicit knowledge basis encoded in human interpretable statements). Proofs of algorithmic predictions require clear causal links and logic statements that unambiguously trace performance from data to decision. Such an examination is possible in AI systems that employ ruled-based or decision-tree models because the rationale is specifically coded and human readable. While the performance of rule-based and decision-tree systems is inferior to that of current machine-learning techniques, these systems are still in use where explicit causality and accountability can be documented (for example, in certain health care, insurance, and public sector applications), demonstrating that, in specific circumstances, explainability is preferred over performance.

## *Validations*

Validations or verifications as explanations are conclusions about the veracity of the AI substantiated by evidence and/or reason. Verification confirms the AI performance against an external measure, standard, factual data, or third-party corroboration.

Feature selection is an explanatory strategy that attempts to reveal the key factors (for example, hyperparameter weights) that had a primary role in the prediction of the algorithm. By isolating or adjusting these elements, it is possible to explain the key components of the decision. There are various feature selection techniques 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. Feature selection provides a verification that certain elements have a primary influence on the prediction thereby explaining why a certain outcome pertains but not another (a contrastive explanation). While such an explanation is used primarily

for designers to adjust their models (that is, it is an error-correction process), allowing users to examine feature selection explanations would provide a justification for why a decision was made and would allow them a basis to challenge that result.

In seeking explanations, people rarely ask for or rely on complete explanations. Rather than reviewing and assessing all the causes (even if provided), people tend to be highly selective. We seek and accept explanations that “satisfice.” Approximation or abstraction are techniques that create a more simplified model to explain the more complex model. Approaches such as model distillation or model-agnostic feature reduction create a simplified presentation of the algorithmic model. This approximation or abstraction may compromise accuracy, but it provides an accessible representation that enhances understandability.

XAI researcher Trevor Darrell believes that “the solution to explainable AI is more AI.”<sup>14</sup> In this approach to explanation, oversight AI are positioned as intermediaries between an AI and its users. These AI have been called “ethical governors,”<sup>15</sup> “flight data recorders,”<sup>16</sup> and, more ominously, “AI guardians.”<sup>17</sup> These examples of intelligent middleware offer the ability to interpolate the values and expectations of third parties, such as research libraries, in the process of deriving an explanation from an AI.

Replication is a recognized verification strategy in many aspects of research. Being able to independently reproduce results in different settings provides evidence of veracity and supports user trust. However, documented problems in reproducing machine-learning research have questioned the generalizability of these approaches and undermined their explanatory capacity. In response, a “Reproducibility Challenge” was created by the International Conference on Learning Representations (ICLR) to validate 2018 and 2019 conference submissions.<sup>18</sup> More rigorous replication through the availability of all necessary components will be important to this type of verification.

## *Authorizations*

Authorizations as explanations are processes, typically involving third parties, which provide an assessment or ratification of the AI. Authorizations might pertain to the AI model, its operation in specific instances, or even the process by which the AI system was created. Examples of authorization include transparency, expertise, due process, litigation, and liability. This section will look at voluntary codes, audit, legislation, and regulation.

Voluntary codes or standards that encourage explanatory capabilities are approaches to explanation supported by the AI industry and professional organizations (for example, Association for Computing Machinery and IEEE). Self-regulation through non-binding codes or standards is a type of governance that some argue is the most effective for rapidly changing technologies. The inflexibility of legislation and regulation might either unnecessarily constrain AI or be ineffective in managing new developments. The “privacy by design” initiative might be a model for something like “explanation by design” whereby prior impact assessment reports, certification requirements, and codes of conduct would provide incentives for more “scrutable algorithms.” Unfortunately, this strategy is undercut by the poor experience with voluntary mechanisms regarding privacy protection.

A commonly recommended approach to AI explanation is third-party auditing. The use of audits or audit principles is widely accepted in a variety of areas. While auditing is typically *ex post*, it can be accomplished at any stage, including design specifications, completed code, operating models, or periodic audits of specific decisions. Auditing for XAI would require trusted auditors, an accepted set of standards to measure against, and the “auditability” of the algorithms or systems. Critics of the audit approach have focused on lack of auditor expertise, algorithmic complexity, and the need for *ex ante* approaches.

The efficacy, and likelihood, of legislation mandating explanatory AI is widely discussed among researchers. While US, and to a lesser extent Canadian, past practice signals a reluctance to legislate in these areas, the EU, France, and the United Kingdom are taking different and more proactive approaches as exemplified by the GDPR. As a result, in Canada and the US, the most common recommendation for AI oversight and authorization is the use of a regulatory agency. Such an agency 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. The breadth and depth of the responsibilities of these agencies varies by those promoting them and by the relevant jurisdiction. Specific suggestions for a public agency include a “neutral data arbiter” with investigative powers like the US Federal Trade Commission, a Food and Drug Administration “for algorithms,” a standing “Commission on Artificial Intelligence,” quasi-governmental agencies such as the Council of Europe, and a hybrid model combining certification and liability. There are few calls for an international regulatory agency despite the global reach for many, if not most, AI systems and services.

### **XAI and Research Libraries**

Algorithmic decision-making is already pervasive in information tools and services acquired, provided, or developed by research libraries. Often the methods and processes of those tools and services are invisible or unacknowledged. If libraries are to trust the quality, value, and credibility of these innovations, it is important that they be explainable.

David Lankes warns of a new digital divide with “a class of people who can use algorithms and a class used by algorithms,”<sup>19</sup> and argues that “librarians need to become well versed in these technologies, and participate in their development, not simply dismiss them or hamper them. We must not only demonstrate flaws where they exist but be ready to offer up solutions. Solutions grounded in our values and in

the communities we serve.”<sup>20</sup> This is echoed by Catherine Coleman in her assertion that librarians can be co-creators of “an intelligent information system that respects the sources, engages critical inquiry, fosters imagination, and supports human learning and knowledge creation.”<sup>21</sup> There are numerous examples, such as Hamlet from MIT, the AI Lab at the University of Rhode Island, and the Stanford Library AI initiative, where machine learning in research libraries is occurring with an emphasis on explainability and accountability.<sup>22</sup>

Developments such as these highlight Chris Bourg’s 2017 suggestion that “we would be wise to start thinking now about machines and algorithms as a new kind of patron.”<sup>23</sup> In doing so, research libraries need to consider not merely how the data can be exposed to algorithmic systems, but the new obligations with respect to data privacy and reuse that may come from this. These implications may extend beyond what is currently considered in research data management protocols.

“**Algorithmic decision-making is already pervasive in information tools and services acquired, provided, or developed by research libraries.**”

An illustration of why research libraries need to accelerate their involvement in AI and XAI arises from a recent breakthrough in the unsupervised text mining of the scientific literature, which demonstrated “that latent knowledge regarding future discoveries is to a large extent embedded in past publications.”<sup>24</sup>

This insight was observed previously during the formative years of Medline<sup>25</sup> and has motivated the current “knowledge validation engine” of Project Aiur from Iris.ai.<sup>26</sup> Each of these projects acknowledges that the structure of scientific communications (for example, the nature of abstracts) enables machine-learning analysis and highlights the need to verify the outcomes by examining the processes. They also emphasize the challenges of explainability when the research literature is being utilized and interpreted using complex and often opaque methods.

It is concerning that these innovations are occurring outside the field of academic librarianship and with little or no involvement of library expertise. If libraries are to shape AI development

“If libraries are to shape AI development and embed values such as explainability in these tools and services, it is essential that the challenges voiced by Lankes, Bourg, and Coleman be acknowledged, accepted, and acted upon.”

and embed values such as explainability in these tools and services, it is essential that the challenges voiced by Lankes, Bourg, and Coleman be acknowledged, accepted, and acted upon. In addition to the focus on innovation in tools and services, academic libraries can further XAI through such avenues as public policy and algorithmic literacy.

### *Public Policy*

A key XAI strategy is to use authorizations, such as legislation, regulation, and audit, as governance methods to support, or even require, explainability. Despite widespread concerns about algorithmic decision-making with respect to bias, discrimination, and unfairness, this is an area that is largely unregulated in Canada and the United States. The AI public policy landscape is nascent. Some have argued for a “regulatory lag” to allow more clarity on how AI will evolve, while others more cynically dismiss all regulations as solving “yesterday’s challenges” and impeding innovation in a globally competitive “AI race.” While premature and reactive regulation is undesirable, neither is an environment where abuses, harms, and predatory practices are allowed to exist.

Research libraries, through organizations such as the Association of Research Libraries and the Canadian Association of Research Libraries, have a strong interest in influencing public policy and have achieved substantial successes in this area, even if only in raising public awareness. While it is argued that blanket AI regulation will be less effective than application-specific regulation (for example,

let those who regulate air travel regulate AI in air travel), there are overarching principles, such as explainability, that cross application boundaries and deserve a different level of attention. Research libraries can be influential in these debates given their expertise in knowledge management and research support, and their concern for the public good.

An interesting example arises in the area of copyright as a result of discussion about the ownership of materials created by an AI. This has led some to argue for the creation of “AI sunshine laws,” which would mirror the idea of the public domain in copyright or patent law. The code and logic of the AI system would, at some point, become public, transparent, and open to scrutiny and reuse. This requirement would position AI within more traditional IP legislation and would extend the notion of public domain into new and likely highly contentious areas.

### *Algorithmic Literacy*

Research libraries, like all libraries, have been active proponents of enhancing literacy, be it traditional reading and writing or more recently digital literacy in all its various forms. While algorithmic literacy can be seen as a subset of digital literacy or computational thinking, it has unique characteristics and applications that deserve specific attention. Just as information literacy provides users with skills and perspectives to assess resources, algorithmic literacy is an explainability strategy allowing users to navigate and utilize algorithmic tools and services.

Calling “algorithmic awareness” a “new competency,” the objective of the 2017 Institute of Museum and Library Services (IMLS) grant proposal from Jason Clark and colleagues at Montana State University is to “find transparency for the invisible logic embedded in our software interactions. Success in this setting would be our community finding new teaching methods and confidence to make this logic visible for our patrons and ourselves.”<sup>27</sup> By linking algorithmic awareness to

“Just as information literacy provides users with skills and perspectives to assess resources, algorithmic literacy is an explainability strategy allowing users to navigate and utilize algorithmic tools and services.”

information and digital literacy, Clark identifies a gap in the Association of College & Research Libraries information literacy framework revealing “a lack of an understanding around the rules that govern our software and shape our digital experiences.”<sup>28</sup>

The anticipated “Algo Report” from Project Information Literacy will present findings from a national study of college students in the US and address “how algorithms affect the information that streams at them constantly throughout the day in order to be truly information literate in the 21st century.”<sup>29</sup>

### **AI-Authorship: An Explainability Sandbox**

An interesting and instructive example of the role of XAI in research libraries arose earlier this year when Springer Nature published an open access book written by AI: *Lithium-Ion Batteries: A Machine-Generated Summary of Current Research*.<sup>30</sup> The author, identified as “Beta Writer,” algorithmically categorized and summarized more than 150 key research publications selected from over 1,000 published from 2016 to 2018, thereby synthesizing a large and complex corpus of the current research literature. The algorithmic processes that created this book, using a combination of various “off the shelf” natural language processing (NLP) tools, included preprocessing the documents to address various linguistic and semantic normalizations; clustering documents by content similarity (that is, the content in the chapters and sections of the book); generating abstracts, summaries, introductions, and conclusions; and finally outputting XML as a completed manuscript.

The details are outlined in a human-written preface and provide an interesting comparison to current cataloging and metadata processes and to accepted scholarly communication practices.<sup>31</sup> Henning

Schoenenberger, director of product data and metadata management at Springer Nature, is clear that the intent of the project is “to initiate a public debate on the opportunities, implications and potential risks of machine-generated content in scholarly publishing.”<sup>32</sup>

Springer has gone to great lengths to document their process, discuss alternative strategies, identify weaknesses and outright failures, and to encourage critical commentary. In many ways they have provided an “explainability sandbox” for scholarly publishing. Determining the value of this and similar books will be achieved in part by interrogating the methods and processes by which they are constructed. In other words, the emerging AI books will need the capacity to explain themselves.

## Conclusion

In his article about stewardship in the “age of algorithms,” Clifford Lynch argues that algorithmic accountability is “the domain of the regulator or social justice advocate, not the archivist.”<sup>33</sup> However, he also notes that “this new world is strange and inhospitable to most traditional archival practice” and that “our thinking about a good deal of the digital world must shift from artifacts requiring mediation and curation, to experiences.”<sup>34</sup> These observations suggest that the role of the archivist (and of research libraries more generally) should indeed include algorithmic accountability because of its centrality to emerging practices.

“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 complexity and opacity of algorithmic decision-making, replete with limitations, outright failures, and dramatic advances, is challenging and changing our notions of information systems and their use. The field of explainable AI has emerged as a set of strategies, techniques, and processes used in a variety of contexts

to facilitate trust and accountability. As key stakeholders in the scholarly communications ecosystem being significantly disrupted by artificial intelligence, 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.

## Endnotes

1. Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: NYU Press, 2018), 1.
2. Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, MA: Harvard University Press, 2015), 8.
3. 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, UK: Routledge, 2011).
4. Pasquale, *Black Box Society*, 18.
5. Mariarosaria Taddeo, “Trusting Digital Technologies Correctly,” *Minds and Machines*, 27 no. 4 (2017): 565–568, <https://doi.org/10.1007/s11023-017-9450-5>.
6. *Explainable Artificial Intelligence (XAI)*, Broad Agency Announcement DARPA-BAA-16-53 (Arlington, VA: DARPA, August 10, 2016), 5, <http://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf>.
7. Matt Turek, “Explainable Artificial Intelligence (XAI),” DARPA, accessed August 30, 2019, <https://www.darpa.mil/program/explainable-artificial-intelligence>.

8. Siddhartha Mukherjee, “The Algorithm Will See You Now,” *New Yorker*, April 3, 2017, 46–53, <https://www.newyorker.com/magazine/2017/04/03/ai-versus-md>.
9. Tania Lombrozo, “The Structure and Function of Explanations,” *Trends in Cognitive Sciences* 10, no. 10 (2006): 464–470, <https://doi.org/10.1016/j.tics.2006.08.004>.
10. Emilee Rader, Kelley Cotter, and Janghee Cho, “Explanations as Mechanisms for Supporting Algorithmic Transparency,” *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York: ACM, 2018), <https://doi.org/10.1145/3173574.3173677>.
11. Dino Pedreschi et al., “Open the Black Box: Data-Driven Explanation of Black Box Decision Systems,” preprint, submitted June 26, 2018, <http://arxiv.org/abs/1806.09936>.
12. Taina Bucher, *If...Then: Algorithmic Power and Politics* (New York: Oxford University Press, 2018).
13. Tim Miller, “Explanation in Artificial Intelligence: Insights from the Social Sciences,” *Artificial Intelligence*, 267 (2019): 1–38, <https://doi.org/10.1016/j.artint.2018.07.007>.
14. Cliff Kuang, “Can A.I. be Taught to Explain Itself?,” *New York Times Magazine*, November 21, 2017, <https://nyti.ms/2hR1S15>.
15. Alan F. T. Winfield and Marina Jirotko, “The Case for an Ethical Black Box,” in *Towards Autonomous Robotic Systems*, ed. Yang Gao, Saber Fallah, Yaochu Jin, and Constantina Lekakou, Lecture Notes in Computer Science 10454 (Cham, Switzerland: Springer Nature, 2017), 262, [https://doi.org/10.1007/978-3-319-64107-2\\_21](https://doi.org/10.1007/978-3-319-64107-2_21).
16. Winfield and Jirotko, 269.
17. Amitai Etzioni and Oren Etzioni, “Keeping AI Legal,” *Vanderbilt Journal of Entertainment & Technology Law* 19, no. 1 (2016): 136, <https://doi.org/10.2139/ssrn.2726612>.

18. Joelle Pineau, “ICLR Reproducibility Challenge: Second Edition, 2019,” Pineau’s website, accessed August 30, 2019, <https://www.cs.mcgill.ca/~jpineau/ICLR2019-ReproducibilityChallenge.html>.
19. Lee Rainie and Janna Anderson, *Code-Dependent: Pros and Cons of the Algorithm Age* (Washington, DC: Pew Research Center, February 2017), [http://www.pewinternet.org/wp-content/uploads/sites/9/2017/02/PI\\_2017.02.08\\_Algorithms\\_FINAL.pdf](http://www.pewinternet.org/wp-content/uploads/sites/9/2017/02/PI_2017.02.08_Algorithms_FINAL.pdf).
20. R. David Lankes, “Decoding AI and Libraries,” Lankes’s website, July 3, 2019, <https://davidlankes.org/decoding-ai-and-libraries/>.
21. Catherine Nicole Coleman, “Artificial Intelligence and the Library of the Future, Revisited,” *Digital Library Blog*, Stanford Libraries, November 3, 2017, <http://library.stanford.edu/blogs/digital-library-blog/2017/11/artificial-intelligence-and-library-future-revisited>.
22. “Artificial Intelligence and Machine Learning in Libraries,” ed. Jason Griffey, *Library Technology Reports* 55, no. 1 (2019), <https://doi.org/10.5860/ltr.55n1>.
23. Chris Bourg, “What Happens to Libraries and Librarians When Machines Can Read All the Books?,” *Feral Librarian*, March 16, 2017, <https://chrisbourg.wordpress.com/2017/03/16/what-happens-to-libraries-and-librarians-when-machines-can-read-all-the-books/>.
24. Vahe Tshitoyan et al., “Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature,” *Nature* 571, no. 7763 (2019): 95–98, <https://doi.org/10.1038/s41586-019-1335-8>.
25. Don R. Swanson, “Medical Literature as a Potential Source of New Knowledge,” *Bulletin of the Medical Library Association* 78, no. 1 (January 1990): 29–37, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC225324/>.
26. Jacobo Elosua, Anita Schjøll Brede, Maria Ritola, and Victor Botev, “Iris.ai’s Project Aiur: An Open, Community-Governed AI Engine for Knowledge Validation,” Iris.ai, May 2018, [https://iris.ai/wp-content/uploads/2018/05/ProjectAiur\\_whitepaper.pdf](https://iris.ai/wp-content/uploads/2018/05/ProjectAiur_whitepaper.pdf).

27. Jason Clark, Lisa Janicke Hinchliffe, and Scott Young, “‘RE:Search’— Unpacking the Algorithms That Shape Our UX,” Institute of Museum and Library Services, 2017, <https://www.imls.gov/sites/default/files/grants/re-72-17-0103-17/proposals/re-72-17-0103-17-full-proposal-documents.pdf>.
28. Jason Clark, “Building Competencies around Algorithmic Awareness” (presentation at Code4Lib, Washington, DC, 2018), Clark’s website, <https://www.lib.montana.edu/~jason/talks/algorithmic-awareness-talk-code4lib2018.pdf>.
29. Alison J. Head and Barbara Fister, “The Algo Report 2019,” Project Information Literacy, March 28, 2019, [https://www.projectinfolit.org/uploads/2/7/5/4/27541717/about\\_pils\\_algo\\_report-2019.pdf](https://www.projectinfolit.org/uploads/2/7/5/4/27541717/about_pils_algo_report-2019.pdf).
30. Beta Writer, *Lithium-Ion Batteries: A Machine-Generated Summary of Current Research* (Heidelberg: Springer, 2019), <https://link.springer.com/book/10.1007/978-3-030-16800-1>.
31. Henning Schoenenberger, Christian Chiarcos, and Niko Schenk, “Preface,” in Beta Writer, *Lithium-Ion Batteries: A Machine-Generated Summary of Current Research* (Heidelberg: Springer, 2019) <https://link.springer.com/content/pdf/10.1007%2F978-3-030-16800-1.pdf>.
32. Lettie Y. Conrad, “The Robots Are Writing: Will Machine-Generated Books Accelerate Our Consumption of Scholarly Literature?” *Scholarly Kitchen*, June 25, 2019, <https://scholarlykitchen.sspnet.org/2019/06/25/the-robots-are-writing-will-machine-generated-books-accelerate-our-consumption-of-scholarly-literature/>.
33. Clifford Lynch, “Stewardship in the ‘Age of Algorithms,’” *First Monday* 22, no. 12 (December 2017), <http://firstmonday.org/ojs/index.php/fm/article/view/8097>.
34. Lynch.

© 2019 Michael Ridley



This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>.

**To cite this article:** Michael Ridley. “Explainable Artificial Intelligence.” *Research Library Issues*, no. 299 (2019): 28–46. <https://doi.org/10.29242/rli.299.3>.

## Research Librarians as Guides and Navigators for AI Policies at Universities

**Geneva Henry**, Dean of Libraries and Academic Innovation, The George Washington University

### Introduction

Artificial intelligence (AI) is a term that is increasingly a part of daily conversations and is being discussed in many different contexts. Commercial applications of the various AI technologies (for example, natural language processing, machine learning, predictive analytics, robotics)<sup>1</sup> are becoming part of mainstream society without people realizing that AI is at work. Searching the internet using popular search engines, for example, can employ deep learning algorithms that continually learn from previous searches. If the same or a very similar search is performed many times, with users consistently selecting the third-ranked return, the search engine will know that the ranking priority should be adjusted so that the most frequently selected result receives a higher ranking.<sup>2</sup> Users generally do not think about how search results are returned; they're just happy to find what it is they're searching for on the first page of the results without having to sift through the 1,000,000+ possible matches that were returned. Even if someone did want to understand how the search results were prioritized, the proprietary nature of commercial products that are using AI to have a competitive advantage in the marketplace makes it impossible to inspect the software behind the decision-making process.

The end-user experience of using AI-enabled products—from search engines, to self-driving cars, to vacuum cleaners that do our housework for us—can be pleasant, but it can also be deceptive. Without visibility into the algorithms that were programmed into the systems by the software developers, the training data sets that were used to enable the algorithms to build a knowledge base, and the ongoing self-improvement processes that drive the decision-making based on

continued use, users are blindly trusting in systems that can have implicit bias programmed into them and limited knowledge that can skew results towards unexpected behaviors.<sup>3</sup>

## **AI and Research Universities in the National and Global Context**

Research universities occupy an interesting space with respect to AI. They, like any other industry, can rely on AI-enabled systems to identify patterns in massive amounts of data about their users (students, faculty, staff, and visitors) and make inferences that provide guidance in better serving these populations as well as predicting future behaviors. As part of their mission to educate, research universities are teaching students about AI, preparing them to develop algorithms and software, along with data analysis; these are key skills that make students computationally adept, thus providing a pipeline of talent for today's workforce.

The other core mission for research universities is research. Many of these institutions are advancing AI technology through ongoing research with significant funding from the federal government. In 2015, the US government invested approximately \$1.1 billion in unclassified research and development (R&D) for AI-related technologies.<sup>4</sup> In fiscal year 2017, that expenditure was more than \$2 billion and for fiscal year 2020 the federal government expects to invest about \$4.9 billion in unclassified AI research.<sup>5</sup> Research expenditures are an important metric in research university rankings, so knowing the focus of federal funding priorities will inevitably lead to more AI-related research initiatives at universities.

The US is not alone in AI R&D investments. In 2016, the US government published a National Artificial Intelligence Research and Development Strategic Plan to establish objectives for federally funded AI research.<sup>6</sup> In 2017, Canada, China, Finland, Japan, Singapore, and the United Arab Emirates released national strategies to promote AI use and development. In 2018, Denmark, the European Union, France,

“With so much funding being funneled towards AI research and the competitive international landscape that has quickly emerged, increased AI research at universities will continue to accelerate.”

India, South Korea, and the United Kingdom released similar strategies.<sup>7</sup> In 2019, US President Donald Trump signed Executive Order 13859, which established the American Artificial Intelligence Initiative to maintain American leadership in AI.<sup>8</sup> The first directive in that Executive Order is to prioritize AI R&D in federal agencies’ annual budgeting and planning process.

The US also updated its AI R&D strategic plan in 2019 to reflect advances that had been made since the plan was first published in 2016.<sup>9</sup> With so much funding being funneled towards AI research and the competitive international landscape that has quickly emerged, increased AI research at universities will continue to accelerate.

With governments around the world launching national strategies in AI, there is now an increased awareness of the need for policies to govern AI technology. Eleonore Pauwels, with the United Nations University Centre for Policy Research, has examined the power of AI converging with other emerging technologies such as cyber- and biotechnologies, affective computing, neurotechnologies, robotics, and automated manufacturing.<sup>10</sup> She explores scenarios in which the technologies, once released from the lab environments where they were created, can create deception, degradation of truth, targeted monitoring of specific populations, exploitation of vulnerabilities in infrastructure, and other actions that have a negative impact on society and governments. Alternatively, the convergence of AI with other technologies has the opportunity to address issues such as famine and disease, healthcare inequalities, military attacks from hostile forces, election fraud, and violent crimes.

Countries have recognized the need to begin developing policies to govern the use of AI technologies, but this is still in an early stage. The

most advanced policies to date are found around the testing and use of autonomous vehicles.<sup>11</sup> Given the significant impacts that AI will have, it will be important to establish policies that can provide a governing framework that promotes ethically responsible behaviors and accountability of AI systems. These policies will need to interoperate at an international level and be mutually recognized by countries impacted by the technologies. A key step in recognizing this may come from a recent alliance of France, Germany, and Japan to jointly fund research into AI that respects privacy and transparency.<sup>12</sup> Their joint call for research proposals states that their goal is to “present the direction of future digital economy and society through technical progress in AI research to strengthen trust, transparency and fairness.”<sup>13</sup> As noted by Brundage and Bryson, “The key question is not whether AI will be governed, but how it is currently being governed, and how that governance might become more informed, integrated, effective, and anticipatory.”<sup>14</sup>

“Given the significant impacts that AI will have, it will be important to establish policies that can provide a governing framework that promotes ethically responsible behaviors and accountability of AI systems.”

While there is a great deal of incentive to pursue federal research funding to continually advance AI, there is a parallel responsibility to ensure that the university is using, teaching, and creating AI technologies responsibly. Issues such as understanding the provenance of the data used that drives automated decisions, being able to examine algorithms for bias, and being attentive to privacy or other ethics concerns are important things to address in all aspects of AI use and development. What measures can be taken to limit the likelihood that a university develops AI technology that leads to nefarious uses? How can campus users of AI-enabled systems guard against decisions that may be guided by biased algorithms? How does instruction that prepares university students for AI work ensure that they are sensitive to issues like bias, data provenance, privacy, and other ethical concerns

that may impact the products they create or decisions they make with AI technology?

An important step in this direction is the establishment of policies related to AI that can inform and guide the university community in all of its areas of work with AI. In addition to setting guidelines for the university, there is also an opportunity for universities to provide leadership and guidance in developing local, national, and international policies for AI technology, given the early state of policies that currently exist around AI.

Competing interests within a university can present challenges in developing policies that align with shared values, ethical responsibility, and respect for individual privacy. The drive to maximize research awards for advancing AI technologies can cause researchers to perceive policies as restrictive barriers to pursuing research opportunities. Staff and administration who desire to learn as much as possible about current as well as prospective students may not favor policies that restrict how data can be used with systems using AI to maximize student success. Requiring faculty who teach AI courses to include ethics, privacy, transparency, and implicit bias training in their curriculum will undoubtedly lead to complaints that there is no room in the curriculum for this added material, let alone the expertise required to teach those subjects. Using AI-empowered systems to assess faculty performance and impact could lead to less-subjective promotion and tenure decisions. If, however, there is a lack of transparency to provide insight into the underlying data and algorithms used, the integrity of the process will be called into question.

### **AI and Research Librarians**

Research librarians, having expertise in information science, are well positioned to navigate the campus landscape and work with stakeholders to form policies that can ensure accountability, transparency, and alignment with ethical values. Long guided by

the American Library Association's Code of Ethics,<sup>15</sup> librarians are sensitive to issues related to information ethics and privacy. They are also aware of information policies more broadly that impact the

“Research librarians, having expertise in information science, are well positioned to navigate the campus landscape and work with stakeholders to form policies that can ensure accountability, transparency, and alignment with ethical values.”

universities and their use of different types of information.

As new forms of information and methods for working with that information have continued to evolve, today's research librarians are instrumental in working with faculty, students, and staff to help with managing information and to provide guidance related to such policies as copyright, intellectual

property, privacy, and ethical use of personal information. Librarians have increasingly become a part of research teams on campus, helping them manage their data and develop consistent, replicable processes for working with their data.<sup>16</sup>

Data is at the core of AI-enabled systems, with data sets used as training for developing a more generalized model that can make decisions on new data.<sup>17</sup> Accustomed to working with faculty, students, and staff, librarians are not only qualified experts in understanding data provenance, but are also trusted professionals who steward information and provide education.<sup>18</sup> Librarians can work to identify areas where policies will be beneficial and bring awareness of existing laws—such as the General Data Protection Regulation (GDPR) passed by the European Union (EU) to protect EU citizens' right to privacy with online information<sup>19</sup>—so that there is consistency with higher-level governance.

By examining some of the issues that exist in AI research, education, and administrative uses at research universities, it is possible to understand the impartial role librarians can have in working with campus stakeholders to develop policies that identify the decisions

that should be permitted and encouraged vs. those that should be managed.<sup>20</sup>

## **A Role for Research Librarians in AI in University Research**

In a recent survey that examined publications from 21 leading scientific conferences in the field of AI in 2018, only 18% of the authors were women. These researchers mostly have PhDs and represent the research underway in AI throughout the world. The US continues to graduate PhDs whose publication rates dominate other countries, with 44% of the 2018 publications produced by scholars who earned their PhD in the US, followed by China at 11%, the United Kingdom at 6%, Germany at 5%, and Canada, France, and Spain each at 4%. Furthermore, the survey also found that the AI talent pool is very mobile, with approximately one third of the researchers employed outside of the country where they received their PhD.<sup>21</sup> When we

“When we look at faculty in AI, both tenured and non-tenured, African American representation is only 1.7%.”

look at faculty in AI, both tenured and non-tenured, African American representation is only 1.7%.<sup>22</sup> This lack of diversity in the university research population is troubling when algorithms and training data sets are being developed and selected

by a mostly homogeneous group of primarily white men. Given the widespread lack of diversity among this population, the normal research review process of peer review will draw from this same population, further exacerbating issues that may be present, such as unconscious bias and data training sets that may be skewed to unfairly represent certain populations.

Research emerging from universities can make its way into industry products and visibility into the algorithms and data training sets can be hidden from the end users. Universities have a responsibility to ensure that the research emerging from the institution is of the highest integrity. Having policies that require accountability of algorithms,

showing not just the algorithm itself and the process followed when using data, but an explanation of the extent to which the data used had an influence on the decision outcome.<sup>23</sup> Data sets that are used to train the systems must also be open to inspection to uncover potential biases and lack of true representation.

Research librarians who are part of AI research teams can be sensitive to the need for well-documented and open systems. Librarians are likely to be aware of how other existing policies will influence outcomes. Librarians working in this space will need a sufficient understanding of algorithms so that they can validate the documentation that explains the algorithm and its intended impacts with the data that is fed into it. To demonstrate replicability and consistency, the algorithms, their explanations, and associated data training sets should be archived by the institution, a role that is well suited to research libraries and a function that data librarians often perform as members of research teams. Making these materials openly available will allow other researchers to replicate the findings and make improvements to further advance research. Concerns researchers may have about others claiming credit by using these algorithms and data can be mitigated by archival restrictions such as embargoes or limited-access restriction if necessary. Establishing institutional policies around documenting algorithms and archiving both the algorithms and training data will benefit the larger research community and can provide a safeguard for the university against the risk of claims associated with harm caused by the use of the technology.

### **A Role for Research Librarians in AI in University Education**

The underrepresentation of women and people of color in AI at the PhD level is a reflection of the underrepresentation that exists at the undergraduate level in students' choice of majors and the courses they take. Undergraduate enrollments in computer science have increased significantly, with a growth of 136% among full-time computer science majors between 2006 and 2015.<sup>24</sup> With this growth, there has not

been an improvement in the representation of women and people of color, which has been historically low. The percentage of information science degrees earned by women has remained steady at about 22%, and between 2011 and 2015 there was an increase from 13.6% to 15.9% of computer science degrees earned by women; this is still a small fraction of overall degrees awarded. The share of computer science, computer engineering, and information science degrees awarded to African Americans decreased between 2009 and 2015 from 8.0% to 6.1% for computer science, 5.8% to 4.9% for computer engineering, and 15.0% to 13.4% for information science.<sup>25</sup> Looking at enrollments in AI courses at Stanford University and the University of California, Berkeley in 2017, both schools reported enrollments in introductory AI courses as approximately 74% male, and introductory machine learning as about 76% male at Stanford and 79% male at Berkeley.<sup>26</sup> Sexist jokes, slide presentations that only show men, and masculine language are some of the reasons that women lack interest in computer science.<sup>27</sup> Understanding that undergraduates in these majors populate the pipeline for graduate students and future PhDs as well as for the workforce looking to hire AI talent, universities must be mindful of the biases that can be formed with this type of white-male-dominated educational environment.

Research librarians at many universities have been actively engaged in the curriculum, partnering with faculty in all disciplines. Engaging librarians in computer and data science courses can help with teaching students about important concepts such as validating information, understanding data provenance, finding appropriate information resources and vetted data to use in their research and experiments, and issues related to privacy and ethical uses of data. Teaching students about good research practices very early—for example, documenting algorithms, questioning data sources, and archiving software and data so that the expected behavior of algorithms can be replicated by others—is also useful information that librarians can teach as part of the curriculum. Librarians are familiar with policies, such as Title IX,<sup>28</sup> related to sexual harassment and inappropriate behavior and

can ensure that students in these courses are aware of their rights and responsibilities with respect to these policies. The risk to a university of being accused of discrimination against or harassment of any student can be very costly to its reputation as well as financially. Worse, sending students into the workforce who have not received an education where diversity, equity, and inclusion are important values that should be fully integrated into all aspects of work can result in practices that perpetuate bias in products that are developed and system behaviors that skew results to perpetuate and exacerbate biases.

A university can develop policies that address known issues in the curriculum and include requirements to teach students involved with software development about concerns such as unconscious bias, privacy, and ethical responsibility when working with personal information. Requiring students to understand data sources and to document software with explanations of expected behaviors when that software interacts with data will benefit society as students enter the workforce. Teaching all students how to assess information validity, regardless of its format, is a critical skill that people should have in our society. Policies can be adopted that lead to computationally and digitally fluent citizens who can assess information produced by software to know if it is valid and who act responsibly in using data. Research librarians who understand these issues and how to work with digital resources are a key resource for developing university policies that help to stem the flow of “fake news” and misinformation that propagates through social media and other online sources. As noted by Matt Chessen in his report, *The MADCOM Future*, “Academia has been essential in developing cybersecurity best practices, and it should do the same in the cognitive security space.”<sup>29</sup> Establishing educational policies that incorporate digital fluency skills in the curriculum can be a good step towards achieving this for society.

## **A Role for Research Librarians in the Use of AI in University Administrative Systems**

The core mission of research universities is to educate and do research. The previous sections addressed these two areas and ways that research librarians can be integrated into the teaching and research activities to help with establishing processes for greater transparency in AI teaching and research. Establishing university policies and practices to support accountability in both research and education can begin addressing issues that AI-empowered systems can cause in society. But the use of AI in higher education is also proliferating in systems being adopted to improve decisions and services for students, faculty, and staff. Universities are investing in enterprise systems that can process massive amounts of data to detect patterns that will help with admissions decisions, student retention, advising, and understanding how students learn (that is, learning analytics).<sup>30</sup> These same approaches can be used to analyze faculty productivity and impact to provide insight into promotion and tenure decisions. While there is clearly a lot of benefit to be gained from these systems, there are challenges with visibility into the algorithms that commercial systems use and a lack of insight into the training data that was used to enable ongoing pattern detection with new data. There are also a number of issues related to privacy and compliance with such regulations as the Family Educational Rights and Privacy Act of 1974 (FERPA).

“ Librarians bring expertise in working with personally identifiable information, data privacy and security, informed consent, and access-controlled data storage. Leveraging this expertise can help universities adopt informed policies regarding the use of AI systems in making decisions that can have a significant impact on students.”

An area where librarians have become increasingly engaged with the use of AI in higher education is the area of learning analytics. With

the ability to collect data on most every aspect of a college student's life, analysis of patterns can provide greater insight and predictions to assist students with their success. To be most effective, student data will need to be shared across university units in order to gain a greater understanding of the student's performance and behaviors. Broadly sharing a student's data across many members of the campus community can increase the risks of violating student privacy and regulatory protections such as FERPA and the Health Insurance Portability and Accountability Act of 1996 (HIPAA). Data security in these environments is also important and can impact how systems work with the data. The Library Integration in Institutional Learning Analytics project (LIILA) has documented their research and analysis of librarian involvement with learning analytics activities on campus.<sup>31</sup> Librarians bring expertise in working with personally identifiable information, data privacy and security, informed consent, and access-controlled data storage. Leveraging this expertise can help universities adopt informed policies regarding the use of AI systems in making decisions that can have a significant impact on students. When black-box algorithms that are not transparent are using increasingly diverse data to make decisions and recommendations, having policies in place to enable more accountability will be important to explain how decisions are being made and confirm that bias is being held in check.

## **Summary**

The computational ability to process massive amounts of data and detect patterns that can refine themselves over time enables a level of intelligence that humans cannot achieve due to cognitive limitations in processing truly large amounts of data and information. As Brundage and Bryson have observed, however, "Artificial intelligence is not necessarily similar or equivalent to human intelligence. In fact, because human intelligence keeps evolving (primarily culturally but even biologically) to meet the requirements of our animal lives and societies, it is unlikely that even if an AI was built to be exactly like human intelligence that it would stay that way for long."<sup>32</sup>

AI manifests itself in multiple ways at our universities, including research, education, and administrative uses of the technology. Universities can have a lasting impact on society by the technology advances that are made and the students who become productive global citizens. Establishing well-informed policies to govern AI at universities can also result in lasting impacts for society with more transparency and accountability of AI algorithms and data.

“**Research universities can and should demonstrate leadership in policy development before policies are developed that hamper future research and the advantages that AI can bring.**”

Research librarians are well positioned to navigate this challenging and evolving landscape. As partners throughout the higher education enterprise, they can provide the necessary guidance that results in sound policies that can be more

widely adopted by and adapted to the larger world. AI governance will continue to grow as the technologies continue to advance and impact our lives in many ways. Research universities can and should demonstrate leadership in policy development before policies are developed that hamper future research and the advantages that AI can bring. Engaging research librarians in the many AI-related activities that take place on the campus leverages their expertise in data privacy, ethics, validating information integrity, data management, heightening awareness of bias in data and algorithms, archiving software and data for use by others to replicate intended outcomes, providing transparency in documenting software behaviors, conforming with existing public policies, and providing access to vetted information sources. Research librarians will be important strategic partners in developing campus policies for AI that lead to greater trust and accountability.

## Endnotes

1. Nisa Malli, Melinda Jacobs, and Sarah Villeneuve, *Intro to AI for Policymakers: Understanding the Shift* (Toronto, ON: Brookfield Institute for Innovation and Entrepreneurship, March 2018), <https://brookfieldinstitute.ca/report/intro-to-ai-for-policymakers>.
2. Yoshua Bengio, “Springtime for AI: The Rise of Deep Learning,” *Scientific American* 314, no. 6 (June 1, 2016): 46–51, <https://doi.org/10.1038/scientificamerican0616-46>.
3. Miles Brundage and Joanna Bryson, “Smart Policies for Artificial Intelligence,” preprint, submitted August 29, 2016, 12, <https://arxiv.org/abs/1608.08196>.
4. National Science and Technology Council and Networking and Information Technology Research and Development Subcommittee, *The National Artificial Intelligence Research and Development Strategic Plan*, Networking and Information Technology Research and Development Program, October 2016, [https://www.nitrd.gov/pubs/national\\_ai\\_rd\\_strategic\\_plan.pdf](https://www.nitrd.gov/pubs/national_ai_rd_strategic_plan.pdf).
5. Chris Cornillie, “Finding Artificial Intelligence Money in the Fiscal 2020 Budget,” *Bloomberg Government*, March 28, 2019, <https://about.bgov.com/news/finding-artificial-intelligence-money-fiscal-2020-budget/>.
6. National Science and Technology Council and Networking and Information Technology Research and Development Subcommittee, *The National Artificial Intelligence Research and Development Strategic Plan*.
7. Tim Dutton, “AI Policy 101: An Introduction to the 10 Key Aspects of AI Policy,” *Medium*, July 5, 2018, <https://medium.com/politics-ai/ai-policy-101-what-you-need-to-know-about-ai-policy-163a2bd68d65>.
8. Exec. Order No. 13,859, 84 Fed. Reg. 3967 (Feb. 11, 2019), <https://www.govinfo.gov/content/pkg/FR-2019-02-14/pdf/2019-02544.pdf>.

9. Select Committee on Artificial Intelligence of the National Science & Technology Council, *The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update*, Networking and Information Technology Research and Development Program, June 2019, <https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf>.
10. Eleonore Pauwels, *The New Geopolitics of Converging Risks: The UN and Prevention in the Era of AI* (New York: United Nations University Centre for Policy Research, April 29, 2019), <https://i.unu.edu/media/cpr.unu.edu/attachment/3472/PauwelsAIGeopolitics.pdf>.
11. *Regulation of Artificial Intelligence in Selected Jurisdictions* (Washington, DC: Law Library of Congress, January 2019), <https://www.loc.gov/law/help/artificial-intelligence/regulation-artificial-intelligence.pdf>.
12. David Matthews, “New Research Alliance Cements Global Split on AI Ethics,” *Times Higher Education* (THE), August 15, 2019, <https://www.timeshighereducation.com/news/new-research-alliance-cements-global-split-ai-ethics>.
13. The French National Research Agency, Deutsche Forschungsgemeinschaft e.V., and Japan Science and Technology Agency, “Trilateral Call for Proposals on Artificial Intelligence (AI),” Deutsche Forschungsgemeinschaft e.V., July 30, 2019, [https://www.dfg.de/download/pdf/foerderung/internationales/dfg\\_jst\\_anr\\_call\\_text\\_2019.pdf](https://www.dfg.de/download/pdf/foerderung/internationales/dfg_jst_anr_call_text_2019.pdf).
14. Brundage and Bryson, “Smart Policies for Artificial Intelligence.”
15. “Code of Ethics of the American Library Association,” American Library Association, amended January 22, 2008, <http://www.ala.org/advocacy/sites/ala.org.advocacy/files/content/proethics/codeofethics/Code%20of%20Ethics%20of%20the%20American%20Library%20Association.pdf>.
16. Joyce M. Ray, ed., *Research Data Management: Practical Strategies for Information Professionals*, Charleston Insights in Library, Archival,

and Information Sciences (West Lafayette, IN: Purdue University Press, 2014).

17. Solon Barocas, Moritz Hardt, and Arvind Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*, incomplete draft, last updated August 21, 2018, <https://fairmlbook.org>.
18. Portland Research Group, Maine State Library, Bruce M. Lockwood, and James Ritter, *Maine State Library: Trusted Professionals Survey 2016*, Library Documents (Augusta, ME: Maine State Library, 2016), [http://digitalmaine.com/msl\\_docs/101](http://digitalmaine.com/msl_docs/101).
19. European Commission, *Communication from the Commission to the European Parliament and the Council: Data Protection Rules as a Trust-Enabler in the EU and beyond—Taking Stock* (Brussels: European Commission, July 24, 2019), [https://doi.org/10.1163/2210-7975\\_HRD-4679-0058](https://doi.org/10.1163/2210-7975_HRD-4679-0058).
20. Brundage and Bryson, “Smart Policies for Artificial Intelligence.”
21. JF Gagne, Grace Kiser, and Yoan Mantha, “Global AI Talent Report 2019,” accessed August 28, 2019, <https://jfgagne.ai/talent-2019/>.
22. Edward C. Dillon Jr., Juan E. Gilbert, Jerlando F. L. Jackson, and LaVar J. Charleston, “The State of African-Americans in Computer Science—The Need to Increase Representation,” *Computing Research News* 27, no. 8 (September 1, 2015), <https://cra.org/crn/2015/09/expanding-the-pipeline-the-state-of-african-americans-in-computer-science-the-need-to-increase-representation/>.
23. Finale Doshi-Velez et al., “Accountability of AI Under the Law: The Role of Explanation,” preprint, submitted November 3, 2017, revised November 21, 2017, <http://arxiv.org/abs/1711.01134>.
24. National Academies of Sciences, Engineering, and Medicine, *Assessing and Responding to the Growth of Computer Science Undergraduate Enrollments* (Washington, DC: The National Academies Press, 2018), <https://doi.org/10.17226/24926>.

25. National Academies of Sciences, Engineering, and Medicine, “Impacts of Enrollment Growth on Diversity in Computing,” in *Assessing and Responding to the Growth of Computer Science Undergraduate Enrollments* (Washington, DC: The National Academies Press, 2018), 230, <https://doi.org/10.17226/24926>.
26. Yoav Shoham et al., *The AI Index 2018 Annual Report* (Stanford, CA: AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, December 2018), <http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf>.
27. Blanca Myers, “Women and Minorities in Tech, by the Numbers,” *Wired*, March 27, 2018, <https://www.wired.com/story/computer-science-graduates-diversity/>.
28. US Department of Education, Office for Civil Rights, *Title IX Resource Guide* (Washington, DC: US Department of Education, Office for Civil Rights, April 2015), <https://www2.ed.gov/about/offices/list/ocr/docs/dcl-title-ix-coordinators-guide-201504.pdf>.
29. Matt Chessen, *The MADCOM Future: How Artificial Intelligence Will Enhance Computational Propaganda, Reprogram Human Culture, and Threaten Democracy...and What Can Be Done About It* (Washington, DC: Atlantic Council, 2017), [https://www.atlanticcouncil.org/wp-content/uploads/2017/09/The\\_MADCOM\\_Future\\_RW\\_0926.pdf](https://www.atlanticcouncil.org/wp-content/uploads/2017/09/The_MADCOM_Future_RW_0926.pdf).
30. Lee Gardner, “How A.I. Is Infiltrating Every Corner of the Campus,” *The Chronicle of Higher Education*, April 8, 2018, <https://www.chronicle.com/article/How-AI-Is-Infiltrating-Every/243022>.
31. Megan Oakleaf, *Library Integration in Institutional Learning Analytics* (Syracuse, NY: Syracuse University, November 15, 2018), <https://er.educause.edu/-/media/files/library/2018/11/liila.pdf>.
32. Brundage and Bryson, “Smart Policies for Artificial Intelligence.”

## Further Reading

- AI Now Institute. *Algorithmic Accountability Policy Toolkit*. Toolkit 01. New York: AI Now Institute, New York University, October 2018. <https://ainowinstitute.org/aap-toolkit.pdf>.
- Broussard, Meredith. *Artificial Unintelligence: How Computers Misunderstand the World*. Reprint edition. Cambridge, MA: The MIT Press, 2019. <https://mitpress.mit.edu/books/artificial-unintelligence>.
- Dutton, Tim, Brent Barron, and Gaga Boskovic. *Building an AI World: Report on National and Regional AI Strategies*. Toronto, ON: Canadian Institute for Advanced Research, December 6, 2018. <https://www.cifar.ca/cifarnews/2018/12/06/building-an-ai-world-report-on-national-and-regional-ai-strategies>.
- Etzioni, Amitai, and Oren Etzioni. “Designing AI Systems That Obey Our Laws and Values.” *Communications of the ACM* 59, no. 9 (September 2016): 29–31. <https://doi.org/10.1145/2955091>.
- Groth, Olaf J., Mark J. Nitzberg, and Stuart J. Russell. “AI Algorithms Need FDA-Style Drug Trials.” *Wired*, August 15, 2019. <https://www.wired.com/story/ai-algorithms-need-drug-trials/>.
- Harwell, Drew. “Defense Department Pledges Billions toward Artificial Intelligence Research.” *The Switch*. *Washington Post*, September 7, 2018. <https://www.washingtonpost.com/technology/2018/09/07/defense-department-pledges-billions-toward-artificial-intelligence-research/>.
- Klutka, Justin, Nathan Ackerly, and Andrew J. Magda. *Artificial Intelligence in Higher Education: Current Uses and Future Applications*. Louisville, KY: Learning House, Wiley Education Services, November 26, 2018. <https://49hk843qjpwu3gfmw73ngy1k-wpengine.netdna-ssl.com/wp-content/uploads/2018/11/201811-AI-in-Higher-Education-TLH.pdf>.

Selbst, Andrew D., and Solon Barocas. “The Intuitive Appeal of Explainable Machines.” *Fordham Law Review* 87 (2018): 1085–1139. <https://doi.org/10.2139/ssrn.3126971>.

Whittaker, Meredith, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. *AI Now Report 2018*. New York: AI Now Institute, New York University, December 2018. [https://ainowinstitute.org/AI\\_Now\\_2018\\_Report.pdf](https://ainowinstitute.org/AI_Now_2018_Report.pdf).

© 2019 Geneva Henry



This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>.

**To cite this article:** Geneva Henry. “Research Librarians as Guides and Navigators for AI Policies at Universities.” *Research Library Issues*, no. 299 (2019): 47–65. <https://doi.org/10.29242/rli.299.4>.

---

**Correction, September 20, 2019:** On page 8, the example of AI in research libraries highlighted in the first full paragraph was identified incorrectly in an earlier version of this publication. The example institution was The University of Oklahoma, not Oklahoma State University.

## Research Library Issues

*Research Library Issues (RLI)* focuses on current and emerging topics that are strategically important to research libraries. The articles explore issues, share information, pose critical questions, and provide examples. Suggestions for potential themes, articles, and authors are welcome. Please [submit suggestions via this online form](#).

ISSN 1947-4911 <https://doi.org/10.29242/rli>

Editor-in-chief: Mary Lee Kennedy

Managing editor: Elizabeth A. Waraksa

Copy editor: Kaylyn Groves

Layout editor: Katie Monroe

© 2019 Association of Research Libraries

ARL policy is to grant blanket permission to reprint as long as full attribution is made. Exceptions to this policy may be noted for certain articles. This is in addition to the rights provided under sections 107 and 108 of the Copyright Act. For more information, contact ARL Publications, [pubs@arl.org](mailto:pubs@arl.org).

Current and back issues are available on the ARL Digital Publications website, [publications.arl.org/rli](http://publications.arl.org/rli). The website is also where you may sign up for alerts to new releases of *Research Library Issues*.

**Association of Research Libraries**

21 Dupont Circle, NW  
Suite 800  
Washington, DC 20036  
T 202.296.2296  
F 202.872.0884

ARL.org  
[pubs@arl.org](mailto:pubs@arl.org)