Applying Concepts of Algorithmic Justice to Reference, Instruction, and Collections Work

Sofia Leung, Teaching and Learning Program Manager and Liaison to Comparative Media Studies/Writing

Michelle Baildon, Collections Strategist for Arts & Humanities and Liaison to the Science, Technology, & Society Program

Nicholas Albaugh, Management and Social Sciences Librarian for Innovation & Entrepreneurship; Economics Librarian

MIT Libraries

August 9, 2019

Introduction

The growth of artificial intelligence (AI), machine learning, and big data present challenges and opportunities to academic and research libraries. These challenges and opportunities are not only operational, but also ethical, social, and political, and they prompt consideration of core professional and organizational values. As we launch the information citizenship program, we can apply the critical approaches that we teach to our daily work. This lens is of core importance as our collections and services pivot towards the computational. A growing number of groups, teams, projects, and initiatives such as the Algorithmic Justice League, Relata, the Ethics & AI Study Group, AI Can't Fix This, and the Schwarzman College of Computing working group on Social Implications and Responsibilities of Computing engage in issues of data and algorithmic justice. We must be ready to engage with our users in this area by building competencies in social, political, and ethical analysis of data, computation, and AI.

As part of the MIT Libraries Library Instruction and Reference Services (LIRS) department's "Summer of Data" initiative, we participated in a project on algorithmic justice. The goal was to explore fundamental concepts and arguments and learn about relevant work being done at MIT to allow us to better support

our liaison communities and participate fully in the information citizenship program. We read or listened to and discussed parts of *Artificial Unintelligence* by Meredith Broussard, Broussard's keynote conversation with Ruha Benjamin from the May 2019 Critical Race and Digital Studies Conference, Benjamin's introduction to *Captivating Technology: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life*, the syllabus developed for the Critical Race and Digital Studies Conference, and MIT Libraries' director Chris Bourg's June 2019 talk for Madroño Consorcio in Madrid.

We apply a number of concepts from Broussard's *Artificial Unintelligence* and Benjamin's introduction to *Captivating Technology* to our work at the MIT Libraries, especially with regard to the kind of future Bourg discusses in her talk. After summarizing some key concepts, we then explore implications for three areas of our work as liaison librarians: reference, instruction, and collection development. Finally, we end with some of the larger implications for the program on information citizenship and the Task Force Report on the Future of Libraries recommendations.

Key Concepts

Artificial Unintelligence

One of the organizing concepts of Broussard's *Artificial Unintelligence* is "*technochauvism*," a "flawed assumption" that "tech is always the solution" (Broussard 7-8). Frequently accompanied by "technolibertarian political values," technochauvinism can be expressed as "an unwavering faith that if the world just used more computers, and used them properly, social problems would disappear and we'd create a digitally enabled utopia." A key element of technochauvinism is the assumption that computers are more objective (and thus in some way "better") than people at solving most problems, including social problems.

Broussard traces technochauvinism to longstanding cultural characteristics of math, engineering, and computer science, including a disregard for societal rules, an overestimation of male mathematical ability, a desire to make science fiction reality, and an idolization of the lone genius (85). She writes, "Disciplines like math, engineering, and computer science pardon a whole host of antisocial behaviors because the perpetrators are geniuses. This attitude forms the philosophical basis of technochauvinism, in which efficient code is prioritized above human interactions" (75). Broussard employs the story of MIT professor Marvin Minsky, who is widely considered to be the creator of artificial intelligence, to demonstrate tech's lack of concern for society and the public good. His encouragement and hiring of students who broke into MIT's mainframe computer to play with it led to the first hackers and the hacker culture that is celebrated at MIT. Broussard also discusses a study by Shane Bench, Heather Lench, and others in which they administered the same math test to men and women and then asked them to assess their own performance. They found that men "consistently thought they scored higher than they actually did" (84). This suggests that "gender gaps in STEM fields are not necessarily the result of women underestimating their abilities, but rather due to men overestimating their abilities" (85).

Popular misconceptions of the reality of artificial intelligence can serve to reinforce technochauvinism. To dispel that confusion, Broussard takes care to distinguish between "general" and "narrow" AI. Pop culture sentient machines a là Hal from *2001: A Space Odyssey* or the Terminator are "general AI," which is currently out of reach of computer science research. Machines cannot "think" like people. Current research in AI is "narrow AI," "a mathematical method for prediction" carried out by "analyzing an existing dataset ["training data"], identifying patterns and probabilities in that dataset, and codifying these patterns and probabilities into a computational construct called a model" (32). Broussard observes that there is a similar conceptual slippage in the term "machine learning." She notes that "computer scientists know that machine 'learning' is more akin to a metaphor . . . it means that the machine can improve at its programmed, routine, and automated tasks. It doesn't mean that the machine acquires knowledge or wisdom or agency" (89).

In contrast to the technochauvinist assumption of technological objectivity, Broussard takes as a first principle of *Artificial Unintelligence* that "[all] data is socially constructed" (18). Created, aggregated, analyzed, and understood by human beings, data necessarily reflects human biases. This is essential to Broussard's demystification of artificial intelligence and machine learning. Rather than autonomous, objective sentience, the "narrow AI" of machine learning is able to offer predictive computation based on socially constructed (and thus biased) training data. As training data is created by people, the same biases inherent in people are inherent in machine learning systems.

Broussard presents an alternative to the technochauvinist idea that automation (including AI) is always the solution: the "human-in-the-loop" system. She observes that "Automation will handle a lot of the mundane work; it won't handle the edge cases. The edge cases require hand curation" (176-177). Because "there are things that a human can see that a machine can't," there are many cases--from automated phone systems to data journalism and beyond-- in which human judgement and interpretation are essential.

Broussard's book cautions us against techno-utopian fantasies; artificial intelligence and machine learning systems are specific and constrained technologies that are socially constructed, and they are created within highly gendered and racialized professional cultures. She writes, "We shouldn't rush to be governed by computational systems designed by people who don't care about or understand the cultural system in which we are all embedded" (83). As a data journalist, Broussard herself builds AI software, and she urges us to understand AI and machine learning as tools with much promise, but that they are suited for particular purposes.

Captivating Technology

In the book's introduction, Ruha Benjamin explains that "*Captivating Technology* examines how the management, control, and 'correction' of poor and racialized people provide the raison d'etre for investing in discriminatory designs." Focusing on computer codes as a social structure, the book analyzes "how technoscience reflects and reproduces social hierarchies, whether wittingly or not."

Echoing and expanding on Broussard, Benjamin points out that the supposed objectivity of technoscientific approaches not only obscures bias; these approaches deceptively "seem to 'fix' the problem of human bias when it comes to a wide range of activities"(Benjamin 2). Benjamin coins the term "New Jim Code" for the "insidious combination of coded bias and imagined objectivity" presented by technological "innovation that enables social containment while appearing fairer than discriminatory practices of a previous era" (3).

The concept of "carcerality" is a key part of the New Jim Code, and the book's approach to carcerality is broad: "Racist and classist forms of social control . . . are not limited to obvious forms of incarceration and punishment; rather, they entail . . . a 'carceral continuum' that scales over prison walls," ranging from "credit-scoring algorithms to workplace monitoring systems" (2). This technologically enabled regime of surveillance is pervasive, yet subtle to the point of near-invisibility. Under the guise of neutrality or even empowerment, these systems instead serve to reinforce racism. As Benjamin argues, the comprehensiveness of the New Jim Code calls for a similarly comprehensive response:

> [T]ruly transformative abolitionist projects must seek an end to carcerality in *all* its forms, from the state-sanctioned exercise of social control a la Big Brother, to everyday forms of surveillance that people engage in as workers, employers, consumers, and neighbors a la little brother. Taken together, such an approach rests upon an expansive understanding of the "carceral" that attends to the institutional *and* imaginative underpinnings of oppressive systems (3).

One aspect of "carcerality" that seems highly relevant to libraries are "the twin processes of classification and containment," which "extend well beyond the domain of policing" and purport to offer "innovative solutions to entrenched social problems" (13). Chapters in the book describe projects that claim to advance "noble aims such as 'health' and 'safety'" while enacting "newfangled forms of social control."[1] Although these forms of control reinforce racial and other hierarchies, "in many cases [they] obscure racist logics and assumptions built into their design, ultimately making it more difficult to challenge and demand accountability" (13).

In a line of thinking parallel to Broussard's technochauvinism, Benjamin states that an "argument of human betterment that surrounds technoscience is not only a shiny veneer that hides complexity and camouflages destructive processes," but also "makes it difficult to recognize, much less intervene in, the deadly status quo" (9). This distortion and obfuscation necessitate a commitment to justice as "an ongoing methodology that can and should be incorporated into design processes" (10). Calls for "inclusion" and "access" are insufficient to this challenge. A true "justice-oriented approach to technoscience . . . starts with questioning breathless claims of techno-utopianism, rethinking what counts as innovation, remaining alert to the ways that race and other hierarchies of difference get embedded in the creation of new designs, and ultimately refashioning the relationship between technology and society by prioritizing justice and equity" (11).

[1] This presents a parallel to the notion of "vocational awe" set forth by Fobazi Ettarh.

Applications

Reference

Discussions of future reference models either whole or partially relying on artificial intelligence and machine learning often argue that these models will be more efficient, freeing up librarians to do other more important tasks. This viewpoint is usually expressed implicitly, but sometimes explicitly as well. This is a prime example of the technochauvinism or technoutopianism discussed earlier.

Broussard uses the example of the dataset of passengers on the *Titanic* to illustrate this point. The dataset includes a number of key data points on each passenger: name, sex, age, passenger class, whether the passenger survived the sinking of the ship, how many parents, children, siblings or spouses they had. This data, minus the survival status of each passenger, is used to train a machine learning system to predict which passengers survived the disaster. The system is able to predict the survival of each passenger with 97 percent accuracy. Broussard sees two major problems with this. First, the data alone ignores the social context of the numbers and not all relevant information is included with the data. The major problem is when you make decisions *based* on this insufficient or limited data. As Broussard puts it "Traveling first class on the *Titanic* meant someone was more likely to survive-but it would be wrong to deploy a model that suggests first-class travelers deserve to survive disasters more than people who travel second or third class" (119). Companies often use machine learning systems in this way as the basis for discriminatory price optimization, charging women, poor people, and customers of color more for products.

The basic problem with this approach in the world of reference services is that any AI or machine learning system that libraries would design and use would rely on training data as the basis for its machine learning model. Any dataset that librarians would pull together would carry with all of the biases and lack of diversity inherent in our profession as it exists today.

Suppose the MIT Libraries were to institute a reference model with AI as exclusive element. What dataset would we use for the training data? Two obvious sources would be the corpus of questions archived in our ask-chat and ask-us systems. This would seemingly be a wonderful and sufficient dataset to train the machine learning system on. Consider the limitations of this dataset. In all likelihood, the dataset does not include enough questions in the areas of women and gender studies, African-American studies, and LGBTQ studies to sufficently train the system, to name just three important disciplines. What about the identities of those who asked and answered the questions in the past? This could bias the system away from the experiences and needs of women, people of color, the LGBTQ community and library staff in these and other groups.

The human-in-the-loop system proposed by Broussard would work well to alleviate the biases and limitations of a reference service model relying entirely on machine learning systems. There are several

forms this could take in the area of reference. One model would be an AI-based system for more basic reference questions with librarians taking on the more complicated, in-depth questions. Another model would have all questions reviewed by a subject librarian after the fact for accuracy, completeness and bias mitigation. A final model would have the machine learning system interacting with a librarian in real time as the questions come in.

Instruction

So often when we go into a classroom or any teaching space, we are faced with limitations such as what the faculty member or instructor has asked us to do or the amount of time with which we have to do it. Acknowledging these limitations, we seek to expand our thinking to what we might expose our students (and faculty) to while still addressing what the faculty/instructor has requested. Much of our instruction is centered on data, whether looking for it, organizing it, describing it, or creating it. Broussard's central concept, that data is socially constructed, is something we should be reminding our students of any time data comes up. In the same way we should for any type of information, it is vital to discuss the context of where we get that data from, who created it, what it was created for, and why.

When we discuss databases, Google searches, and any of our collections, we have a responsibility as people working higher education to address the biases and limitations of all of those tools and resources. As we know, a good research article also includes the limitations of the study so that readers understand some of the context within which the study was done. As Broussard says, "Computer systems are proxies for the people who made them," (67). So if we do not mention that all of these systems and collections are made by other human beings and that those systems and collections are replicating the same biases that their creators have than we are missing teaching opportunities with our students. By not mentioning those things, we are making a choice not to counteract algorithm injustice, but to continue to perpetrate it. Of course, it is up to the students to decide what to do with that knowledge, but it remains our responsibility to illuminate the limitations of our knowledge organization systems that could easily be ignored.

As Broussard's book demonstrates, stories of where tech went wrong can be powerful in helping others understand the larger implications of creating and releasing new technology in the world can be. Mathematicians and scientists have a history of making poor choices that only benefits them and other wealthy white men. Broussard writes that the progress of the Industrial Revolution hit up against a limited supply of trained mathematicans because in the nineteenth century, only men were considered eligible for this work (77). Most women did not have the mathematical education to perform the calculations expected of human computers. Slavery was not abolished throughout the country until June 19, 1865, now known as Juneteenth. Enslaved people were certainly not given the type of liberty and education necessary for this work. As Broussard points out, nineteenth-century mathematicans could have advocated and enacted social change to "develop the existing workforce by allowing all the people who weren't elite white men greater access to education and train these workers for jobs" (78). Instead, they chose to build machines to do the work instead. As a society, we are still dealing with the aftermath of that choice and the choices that followed - the gender pay gap, the lack of diversity in tech (and other

professions, including librarianship), and the invisible labor women and people of color often must engage in. We have to be careful that we do not encourage students to pursue the same misguided approach by neglecting to mention the harmful belief that technology, particularly AI and machine learning, is the only solution to humanity's problems.

We should also rethink the way we evaluate our impact in ways that move beyond our current collection of instruction statistics and the data-driven, quantitative approach it privileges. How might we approach our instruction assessment in a justice-oriented way? What metrics beyond the number of workshops taught or the number of students served can we measure?

Collections

In a June 2019 talk for the Madroño Consorcio in Madrid, Chris Bourg shared "some of what we are doing at MIT to reimagine what a research library can and should be and do in a computational age." Bourg described the MIT Libraries working to "align with MIT's core missions: open scholarship, and computational and algorithmic access to collections" with a vision "to be an open, interactive and computational library." Under this vision, the library must be "accessible by machines and algorithms, not just by people. In a computational age, we have to realize that humans are not our only patrons."

How might the staff of MIT Libraries manifest this vision in the realm of collections? Bourg suggests that we can "do what libraries have always done and be a centralized, accessible, and inclusive resource for our communities," including "maintain[ing] online libraries of training data and basic algorithms that students can use and modify as they learn." If we advance collection development practice by including training data and algorithms under our purview, then the Libraries might advance data and algorithmic justice by prioritizing the acquisition of data sets and algorithms that make possible the work of social justice, or that tend towards more just outcomes. To ensure a sufficient critical understanding of the implications of this collection development, we might seek consultation from MIT researchers doing relevant work, or even undertake collaboration to collect their data sets and algorithms, as appropriate, for reuse by other patrons.

In her talk, Bourg argues that:

> the most important thing that libraries can do is work to ensure that the knowledge and research products we already collect, curate, and disseminate are openly available and that the scholarly record is as diverse and as inclusive as possible. Because it is the combination of truly open access to lots and lots of content – text, data, code, images – analyzed with powerful computational tools and methods where really interesting things can happen. . . . Certainly, the choice of topics and problems, the interpretation and application of results requires human imagination – but machine learning tools can speed up the process and, when combined with open access, equalize the ability of people to make use of the knowledge we have already accumulated.

The massive challenge in this formulation is in ensuring that the scholarly (and cultural and historical) records are, in fact, diverse and inclusive. Digitizing and making MIT collections as they currently exist openly accessible at a large scale would inevitably serve to reproduce inequity. Our collections reflect the biases of the Institute, scientific and technical disciplines, and the global scholarly communication system, all of which are shaped by structural racism, patriarchy, and the legacy of colonialism (Baildon, et al.). The collections of the MIT Libraries depict specific viewpoints and experiences and also represent the methods of historically biased disciplines.

To counteract this bias, the MIT Libraries could begin a robust program of collecting perspectives from the margins and from around the world, with an eye towards making them openly computationally available. Such an approach would require the "hand curation" described by Broussard for "edge cases"--which marginalized perspectives surely are. We must also consider that there are risks in representing marginalized communities computationally, exposing data to potential carceral uses. These new approaches to collection development will have to be done in partnership with those communities.

What about the idea of using AI approaches to build collections? We can imagine a machine-learning approach to collection development that analyzes metadata about past monograph purchases to make new selections. This could serve as a supplement to our already existing automated (and algorithmic) approach to collection development--our approval plan for English-language monographs with GOBI Library Solutions--with the training data consisting of metadata about firm orders from GOBI and other vendors. Of course, the training data in this scenario would reflect the many biases of our existing collection development process. These range from the biases of selectors to those of Library of Congress classification to fundamental inequities that facilitate or hinder ready access to publication for certain authors, on certain topics. Perhaps the best approach we have to building collections founded on social justice is to continue and extend an already existing approach: to optimize the current automated, algorithmic system of approval plans for selection of mainstream, commercial material. We might then redeploy selectors' effort and attention towards the "hand curation" required to build diverse, inclusive, and equitable collections, and towards global, systemic efforts to help make more equitable publication possible.

By shedding light on carceral systems that employ harmful uses of people's data, Ruha Benjamin prompts us to move beyond a construct of "data privacy" that emphasizes the individual patron in favor of a broader framework of racialized surveillance. Expanding our view beyond local approaches and fixes that protect patron privacy, can we use our expertise and influence to safeguard against carcerality in electronic information systems?

Conclusion

The Program on Information Citizenship (PIC) is meant to prepare and support MIT community members to be critical creators, consumers, and influencers of the information ecosystem towards the goal of a more just world. The principles we teach MIT students under the PIC program must also stand as guiding principles for our work more generally. As the program develops, the Libraries will require a

justice-oriented, holistic understanding of the limitations and risks of artificial intelligence and machine learning as they apply to our work. The more we rely on these tools, the more we need to be wary of falling for the allure of techno-utopia. We need to ask ourselves, is A.I. the right solution to this problem? We also need to be clear with ourselves, our community of users, and the general public about the reality of what is represented when we make everything we "collect, curate, and disseminate" as open as possible. Who is doing the collecting, curating, and disseminating? What are the biases inherent in that work? As we continue to build our knowledge systems, we need to examine whose research we consider scholarly, valuable, and worth curating. Who continues to be left out of that collection, curation, and dissemination?

Just as we must be critical of ways the Libraries collects, curates, and disseminates data *for* our community of users, we must also be critical of how we create and collect data *about* our community of users. Although this white paper does not explore all the ways in which libraries and archives collect data, we think it is critical to highlight one recent example. In a public discussion at MIT Libraries celebrating Juneteenth (2019), guest speaker Jarrett Drake described how security processes for visitors to archives mirror those of visitors to incarcerated people. He asked, why do we need to collect personal information about our users? What do we use that information for? How could this information be misused if we do not responsibly dispose of it? In all of our practices, the MIT Libraries also must be critical creators, consumers, and influencers of the information ecosystem towards the goal of a more just world.

References

Baildon, Michelle, Dana Hamlin, Czeslaw Jankowski, Rhonda Kauffman, Julia Lanigan, Michelle Miller, Jessica Venlet, and Ann Marie Willer (2017). *Creating a Social Justice Mindset: Diversity, Inclusion, and Social Justice in the Collections Directorate of the MIT Libraries.* Retrieved August 9, 2019: http://hdl.handle.net/1721.1/108771.

Benjamin, R. (2019). *Captivating technology: Race, carceral technoscience, and liberatory imagination in everyday life.* Durham, NC: Duke University Press.

Bourg, C. (2019, July 3). Libraries in a computational age. Retrieved July 23, 2019, from Feral Librarian website: https://chrisbourg.wordpress.com/2019/07/03/libraries-in-a-computational-age/

Broussard, M. (2018). *Artificial intelligence: How computers misunderstand the world.* Cambridge, MA: MIT Press.

Drake, J., & Leung, S. (2019, June). *A Juneteenth Conversation with Jarrett Drake*. Presented at the MIT Libraries, Cambridge, MA. Retrieved from https://calendar.mit.edu/event/a_discussion_with_jarrett_drake#.XU3U_VB7kUo

Ettarh, Fobazi (2018, January). "Vocational Awe and Librarianship: The Lies We Tell Ourselves." *In the Library With the Lead Pipe.*

Kido Lopez, L., & Land, Jackie. (2019, April 18). Critical Race & Digital Studies Syllabus. Retrieved July 23, 2019, from Center for Critical Race and Digital Studies website: https://criticalracedigitalstudies.com/syllabus/

MIT Ad Hoc Task Force on the Future of Libraries. (2019). *Institute-wide Task Force on the Future of Libraries*. Retrieved from https://mitl.pubpub.org/pub/future-of-libraries