Data Science Intelligence: Mitigating Public Value Failures Using *PAIR* Principles

Thema Monroe-White
Berry College
tmonroewhite@berry.edu

Brandeis Marshall
The DataedX Group, LLC
brandeis@dataedx.com

Abstract

In this article, we introduce the term "data science intelligence" as the verified and validated qualitative and quantitative outcomes of the data science workflow. This framing marries the disciplines of science policy and data science in order to empirically ground a way forward for mitigating public value failures resulting from the implementation and use of data science algorithms and practices. After identifying the public value failures in the data science ecosystem, we discuss two public value failures which offer significant challenges and opportunities for data scientists and the organizations they serve. Finally, we pose the Participation, Access, Inclusion and Representation (PAIR) principles framework for organizations seeking to minimize the impacts of these failures via the creation of a taxonomy capable of deploying data science that reflects the values of the communities they aim to serve. Preliminary quantitative outcomes are shared while future work will engage its qualitative aspects.

Keywords

Data science, data science intelligence, public values, marginalized groups

Introduction

Recent McKinsey Global Institute (Henke et al. 2016) and Business-Higher Education Forum (Price Waterhouse Coopers 2017) reports accentuate the critical need for data science practitioners in light of very short supply. Building capacity in data science and its sub-fields, like machine learning (ML) and artificial intelligence (AI), is vital to addressing high demands for data-related skills within academia, industry and government (Chui et al. 2017). However, data science as a field has several significant compounding challenges, mostly related to its newness as a formalized discipline (National Academies of Sciences, Engineering, and Medicine 2018). The absence of a standardized data science curriculum results in ill-defined educational pathways for those seeking entry into data science careers (Berman et al 2018; Marshall 2017). There is general consensus that data science is an inherently interdisciplinary field, residing at the intersection of mathematics, statistics, computer science, information systems and business disciplines (Nazrul 2018; Song and Zhu 2016); however, given that "data science has no natural home (Finzer 2013, p. 1)" strategies for how best to train future data science professionals remain unclear. Furthermore, the scarcity of talent has resulted in a flurry of bootcamp style data science courses and MOOCs aiming to fill the void and capitalize on industry demand (Kross et al. 2017). Despite the dramatic increase in course offerings, women and marginalized groups (i.e., those who identify as African-American/Black, Latinx or American Indian/Alaskan) remain underrepresented in data science. Recent reports found that just 18% of data science professionals were women (Bayern 2019; Harnham 2019) and of students enrolled part-time in a 10-12-week data science course, just 4% identified as Black and approximately 8% identified as Latinx (General Assembly 2019). Given that data science is a nascent discipline, we have yet to uncover the reasons why individuals enter, stay and/or exit the field of data science, only time will tell. However, given systemic historical social and structural barriers, this pattern will only be mitigated through the concerted efforts from academia, government and industry. This research unpacks the challenges and opportunities in data science by adopting a framework developed to identify 'public value failures' in the current data science educational landscape. The public value failure framework (Bozeman 2002)helps answer the questions "who benefits?" and "who is harmed?" Likewise, this framing helps us to remain conscious of the

ways in which formal and informal science and technology institutions (North 1991) benefit some groups and disadvantage others (Cozzens 2007) while ensuring that public goals are achieved.

Conceptualization of Data Science Intelligence

The varied framing of 'what is data science?' necessitates the establishment of a singleton definition that serves as a consistent point of reference (Figure 1a.). We use the following definition (Marshall and Geier 2019):

"Data science is an ecosystem dedicated to the systematic collection, management, analysis, visualization, explanation, and preservation of both structured and unstructured data."

Through an evolution of scientific methods and processes, the field of data science intends to iteratively extract impactful knowledge and insights to better the human condition. An overview of the data science workflow elements is presented in Figure 1b (Azam 2014). There are five elements: 1. Data acquisition and cleanup with a focus on data sourcing, collection, and (re)formatting. 2. Storage and management considerations deal with how to organize data collected for effective handling of complex requests. 3. Data analysis, AI/ML algorithms and processes are leveraged to transition raw data into actionable insights or information. 4. Data visualization focuses on presenting findings in a visual form to help data practitioners determine if their analyses make sense. 5. Data communication and storytelling address the data practitioner's journaling of these phases alongside the dissemination of insights in an accessible manner for non-data professionals. The output of one phase feeds as input into the next phase in hopes of achieving a comprehensive understanding of the data. This understanding shifts as new analyses, visualizations and narratives are shared with human stakeholders. With each phase, there are a suite of technologies and common practices intended to support comprehension.

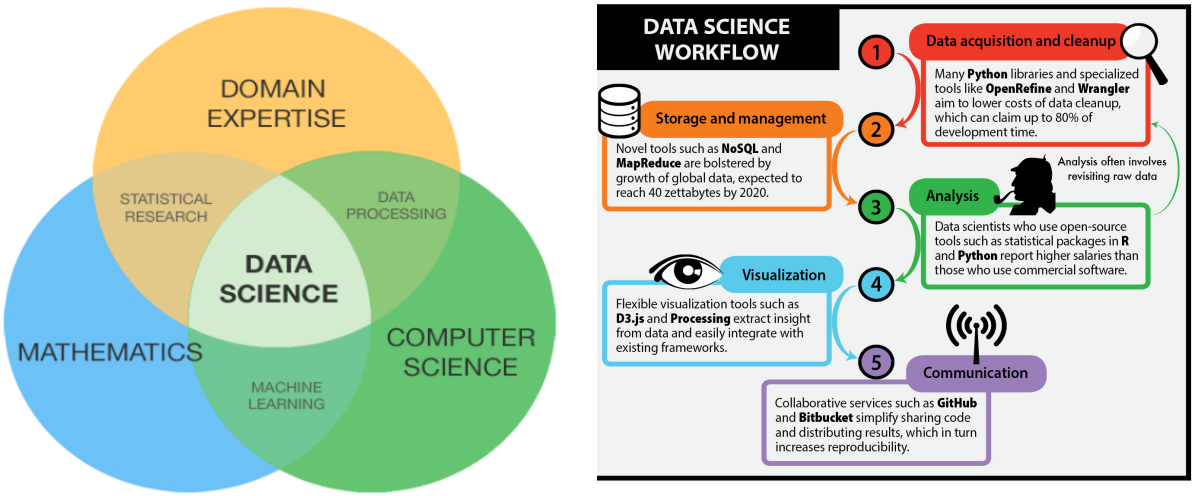

Figure 1. (a) Disciplines that comprise data science (Nazrul 2018) Figure 1. (b) Data science workflow (Azam 2014)

Data science permeates all aspects of our daily lives (MacPhail 2015); however, the policies capable of framing our data science ethics, governance and management strategies are limited or non-existent (McNeeley and Hahm 2017). Furthermore, although situations in which major scientific or technological innovations have outpaced the rate of government regulation are not new (e.g., self-driving cars), the scope and scale of the "known" and "unknown" unknowns of the present data revolution (Dietterich 2017), requires our collective attention and immediate action. As with other socio-technical revolutions (see industrial, post-industrial, information etc.) the so called "fourth industrial revolution (Schwab 2017)" has led to inequalities and disparities in the access, use and empowerment of data science resources (Kneese et al. 2014). These disparities are exacerbated by the fact that data science is both ubiquitous and invisible,

simultaneously and interchangeably benefitting and harming segments of the population with or without their knowledge (Croll 2012; Daniel et al. 2019; O'neil 2016).

The futility of attempts to "opt-out" of data science including the data, code, visualizations, insights and recommendations, necessitates that we evaluate data science products with respect to social implications, as opposed to the scientific or technological advances alone. This evaluative approach is what we have coined "data science intelligence." Leveraging insights from the science policy literature we refine our operationalization of data science intelligence by adopting the public value failures framework (Bozeman 2002, 2003) to explicitly identify the harms (both potential and realized) resulting from information asymmetries in the field of data science at different stages of this workflow.

"Data science intelligence is the verified and validated qualitative and quantitative outcome of the data science workflow."

Our conceptualization of data science intelligence derives from the interdisciplinary fields of data science and science policy. Each of these disciplines is primarily "applied" in focus, emerging out of a need to investigate the role of science and technology (inclusive of the opportunities and challenges) facing actors in both the public and private sectors. The verification and validation of the data science workflow helps to ensure the appropriate access to and use of data science products and processes by the public; and the focus on both quantitative as well as qualitative evidence reinforces the importance of diverse ways of knowing (see the cognitive diversity and cognitive justice literature) when measuring the outcomes and impacts of data science process flows. This combination of insights, in our opinion, is what will advance data science as a discipline while simultaneously addressing the litany of social challenges that are rapidly emerging from the data science profession.

Inherent in this definition of data science intelligence, is the assumption that the ultimate goal of the data science community is to produce desirable outcomes for society. Science is a public institution and has historically served as the principal means by which societies have sought to advance economic growth, security, health and overall welfare of its citizens (McNie et al. 2016). Vannebar Bush, summarizes this view in his groundbreaking presidential report: "[s]ince health, well-being, and security are proper concerns of Government, scientific progress is, and must be, of vital interest to Government. Without scientific progress the national health would deteriorate; without scientific progress we could not hope for improvement in our standard of living or for an increased number of jobs for our citizens; and without scientific progress we could not have maintained our liberties against tyranny." (Bush 1945). Therefore, the relationship between public values and scientific progress has always been bi-directional. The U.S. taxpayers invest in research which in turn produces the revolutionary advancements (e.g., the internet, GPS technologies, HIV/AIDS treatments etc.) that benefit the public. Our desire to adopt the public value failures framework (see explanation below) simply concretizes this understanding and provides a solid basis by which to unpack some of the challenges facing data science as a new and rapidly expanding scientific discipline and applying insights from science policy to the data science intelligence context.

Public Values and Public Value Failures

Public values are defined as "those providing normative consensus about the rights, benefits, and prerogatives to which citizens should (and should not) be entitled; the obligation of citizens to society, the state and one another; and the principles on which governments and policies should be based (Bozeman 2007, p. 17)." A public value failure is "a failure in a society's provision of a public value" (ibid., p. 16) and is accomplished when "neither the market nor public sector provides goods and services required to achieve public values" (ibid., p. 144). Therefore, public value failure theory is distinct from both economic traditional market-failure theory, in which problems are defined by their lack of efficiency; and non-market (i.e. government) failures in which problems are defined with respect to distributional effects (Wolf 1979). Instead, public value failures are defined with respect to their being essential to the human condition.

We define 'the public' as those impacted by the products or outputs of data science, while its 'citizens' are both the producers and consumers of data science including managers, programmers, researchers, and all those who contribute to the production, dissemination and validation of data science products. Table 1

provides a summary of the nine essential public value criteria, how they are defined (Bozeman 2002; Bozeman and Johnson 2015), along with examples of public value failures in the context of data science.

Public Value Criterion - Definition	**Examples of Public Value Failures in Data Science**
PV1: Mechanism for values articulation and aggregation - When political process and social cohesion are limited such that the communication and processing of public values is ineffective.	Public discourse surrounding investments in data science technologies is limited or non-existent (e.g., deepfake detection and regulation mechanisms) potentially leading to increased "tech harms" (Hill Happenings, 2019)
PV2: Legitimate monopolies - When private sector provision of goods would be better administered by government monopoly control.	Data privacy policies vary widely within the private sector. Standard data privacy protection mechanisms should be enforced by government mandated regulations. (Cath 2018)
PV3: Imperfect public information - Lack of transparency leading citizens to making decisions based on incomplete or inaccurate data.	Government agencies adopt privately developed software whose opaque algorithms inhibit transparency and are implemented without the public inquiry or comment. (Buolamwini and Gebru 2018)
PV4: Distribution of benefits - Benefit hoarding such that goods and services are not distributed equally.	The data science educational model benefits those with temporal and financial resources to join predominantly non-accredited, costly, immersive, short courses.
PV5: Provider availability - Scarcity of providers when an essential good or service is needed.	The quantity and quality of data science instructors capable of disseminating content equitably is constrained. (Mantha and Hudson 2018; Noren et al. 2019)
PV6: Time horizon - When short-term market success can lead to long term public failure.	An emphasis on data science adoption, including ML and AI algorithms, as opposed to its social implications may result in deleterious long-term effects. (Kleinberg et al 2016)
PV7: Sustainability vs. conservation of resources - Distinct, valued common resources should be recognized as such as opposed to being substitutable.	Common digital resources require protection. These public goods can become corrupted (i.e., social bots) and harmful for public use (i.e., dis-information) (Ferrara et al. 2016; Vosoughi et al. 2018).
PV8: Ensure subsistence and human dignity - When nations cannot provide basic dignity and subsistence to its citizens, threatening the interests of individuals and the nation.	Data science driven tech harms (e.g., deep fakes) limit citizens' basic human rights (i.e., the right of every individual to control use of their name, image, likeness, or other identifying characteristic).
PV9: Progressive opportunity - Policies that fail to address "structural inequalities" often based on historical differences with regard to access to resources for disadvantaged groups.	Algorithms encode and reinforce existing institutional biases limiting access to quality public education, healthcare, or legal representation by disproportionately disadvantaging marginalized communities.

Table 1. Summary of Public Value Failures 1-9 in Data Science

This public value framing exercise led us to focus our attention on two of the most pressing (with respect to impact) and promising (with respect to mitigation) public value failures: benefit hoarding (PV4) and provider availability (PV5). Addressing these two primary failures, both necessitates and facilitates the systematic treatment of the others.

The Problem: Benefit Hoarding and Provider Availability

The producers of data science tools, algorithms and insights do not currently reflect the demographic, geographic and cognitive diversity of the communities they aim to serve (Harnham Report 2019) nor are these scholars devising the current slate of AI research innovations (Noren et al 2019). There are several reasons for this mismatch, but one of the first relates to the unequal distribution of benefits or benefit hoarding (PV4). It is hard to be "intelligent" about data science if you are not at the table when it is being designed and disseminated. Similarly, if you are unaware that your likeness is being used without your consent, or if you accept the narrative that data science (incl. artificial intelligence, machine learning etc.) are inherently "good" and "objective"; you may miss the often subtle ways in which you are being socially engineered to not question or challenge the status quo (i.e., the absence of protective mechanisms and the growing "tech harms" landscape, Duff 2005). The second reason for the lack of representation in data science has to do with an overall scarcity of providers (PV5). The lack of availability of well trained and experienced data science expertise has to do (in part) with the nascent nature of the field, as well as the poor conceptualization of what makes a "well-trained" data scientist (Kross et al. 2017).

However, the dearth of data science talent has even more to do with the public value failure of benefit hoarding and the socially embedded nature of the data science community (Granovetter 1973; Uzzi 1997; Wellman 2001). Given that a demographically narrow segment of the population dominates the field of data science producers, asking them to prioritize diversity while tapping into such a shallow network to begin with, is particularly challenging. The relative homogeneity of the data science producer community means that they are more likely to search within their existing social network or talent sphere and subsequently hire producers who are demographically similar to themselves, further compounding the problem of lack of diversity and inclusion with each iteration (Harnham Report 2019).

A Proposed Framework: PAIR Principles

A society and its educational structures have a symbiotic relationship, its gaps, e.g., public value failures, can be isolated and mitigated. We focus on PV4: distribution of benefits and PV5: provider availability, due to their positive cascading effect on all of the other public values. The ways, means and personnel distributing data knowledge impacts the actual and perceived public value of that knowledge. We employ data science intelligence through the quantification lens of race/ethnicity and gender disparities in data science by mapping them to established approaches in the science and innovation policy literature.

The Participation, Access, Inclusion and Representation (PAIR) principles framework attempts to address public value failures in data science by accurately portraying what scholars and practitioners want to do, advance public welfare by computationally minimizing discrimination. To accomplish this, we must start by ensuring that those responsible for learning, teaching and doing data science represent the communities they aim to serve.

The PAIR principles model addresses the two public values PV4: distribution of benefits and PV5: provider availability by addressing the diversity, equity and inclusion gaps that lie at the heart of both these public value failures. It is time to shift past the neutral state of acknowledging these inequalities and enact change across every facet of an organization. This growth must happen in two arenas simultaneously — in staffing and technology. First, we must conduct an audit of existing organizational culture. Second, we must set a strategic plan for embedding and sustaining a growth culture. Lastly, we must repeat the first and second step every five to six months. Motivating questions for the audit and implementation phases include:

- *Participation*: Who is participating in data innovations? What are the team dynamics? Who is impacted by the data? Who benefits by this participation?
- *Access*: Who has access to the data? What is the access to systems containing data? What is the availability of tools to create the innovations? Where are the outcomes disseminated for public consumption? What are the mechanisms that promote or thwart access?
- *Inclusion*: What are the intentional and sustainable actions taken to involve and engage those from marginalized communities? Have barriers to inclusion been minimized or removed? what evidence can be provided to show proof of inclusion activities?

- *Representation*: What is the level of representation of those from all demographics within the data, systems, frameworks, etc.? What type of representation is showcased, e.g., promoting or degrading dataset case studies, gender equity, etc.? Which individuals are credited with contributions and the work associated with the data?

Methods and Results

We demonstrate the implementation of our PAIR principles by seeking to answer two of its quantitative aspects: who is participating in data innovations (a Participation principle) and what is the level of representation of those from all demographics within the data, systems, frameworks, etc. (a Representation principle). While both questions overlap both PV4 and PV5, we classify the Participation principle as being more aligned with PV4 given its focus on receiving services, e.g., data science learners, while the Representation principle is associated with PV5 due to its delivery of services, e.g., data science instructors.

Using the state of Georgia as our use case, we compare U.S. national trends in data science instructional offerings (i.e., data science degrees, programs and certificates) by discipline (i.e., Business, Computer Science/Engineering and Math/Statistics) and map these patterns to the Georgia colleges and universities found in the Department of Education's Integrated Postsecondary Education Data System (IPEDS). The latest available IPEDS data (2016 - 2017) were used to 1) identify institutions with one or more of the disciplines known for providing data science instructional offerings and 2) describe the race, ethnicity and gender of both providers (i.e., tenured or tenure-track faculty) and consumers (e.g., students) of data science instructional offerings. Student race/ethnicity and gender by major field of study and tenured or tenure track faculty were used as proxy variables for calculating the relative diversity of instructors and potential data science students within the State. Georgia was chosen as our use case because of its relatively large representation of marginalized groups (i.e., African-American/Black, Latinx or American Indian/Alaskan) compared to the U.S. national average (42.7% vs. 33%) (U.S. Census Bureau 2018); the large and growing technology and innovation sector (Arend 2019). and the wide variety of higher education institution (HEI) types including historically black colleges and universities (HBCUs), liberal arts colleges and research-intensive institutions (U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System 2016). These findings illustrate the overall underrepresentation of both providers and consumers of data science from marginalized communities of color (African-American/Black, Latinx or American Indian/Alaskan) in Georgia, further justifying our focus on benefit hoarding (PV4) and provider availability (PV5) (Table 2). This pattern is also not surprising given persistent disparities in the faculty/student diversity across the U.S. (Leslie and Richard 2019).

Race & Ethnicity (Total %)	**State of Georgia (GA) population**	**GA Instructors: Tenured or Tenure Track Faculty** *(total%: female%; male%)*	**GA students in Data Science disciplines** *(total%: female%, male%)*
African-American / Black	32.4%	10.87%: (5.87%; 5.00%)	26.23% (14.79%; 11.44%)
Latinx	9.8%	2.67%: (1.08%; 1.59%)	6.86% (2.95%; 3.91%)
Asian	3.8%	9.99%: (3.28%; 6.72%)	7.12% (2.82%; 4.30%)
American Indian / Alaskan Native	0.5%	0.21% (0.09%; 0.12%)	0.26% (0.13%; 0.13%)
Native Hawaiian / Pacific Islander	0.1%	0.12% (0.05%; 0.07%)	0.12% (.05%; .07%)
White	52.4%	68.65% (28.58%; 40.06%)	77.75% (29.99%; 47.76%)

Table 2. Race/Ethnicity and Gender composition of HEI providers and consumers of data science in Georgia.

The results above illustrate the severity of the provider and consumer availability problem in data science. The lack of diversity and inclusion in data science is more critical than a mere legal obligation or moral imperative. It has to do with the fact that as a general rule of thumb, one community cannot be expected to effectively represent the interests and needs of another; and that attempts to do so without adequately incorporating the voice, perspectives and values of the communities it aims to serve will simply perpetuate this failure to provide essential public values (Deng et al. 2016). Thus, the lack of diversity, equity and inclusion in data science creates a pernicious feedback loop in which benefit hoarding produces a scarcity of providers, and continued benefit hoarding by a demographically narrow segment of the population.

Discussion

The insights generated from this public value failure framing as well as the results of our analysis of the data science instructional offerings in the state of Georgia provide shine a light on the race, ethnicity and gender disparities in the data science ecosystem. The PAIR (participation, access, inclusion and representation) principles taxonomy helps pave a way forward for producers and consumers of data science content (i.e., platforms, tools, and analysis) to assist in mitigating these disparities.

Conclusion

Data science is leaving behind marginalized communities. The harms associated with technologies that advantage one group over another are well documented; however, the pervasive long-term effects of differences in the production, access and use of data science are as yet unknown given the nascency of the field. Historically, these technological gaps have exacerbated pre-existing economic, health and other structural inequalities, worsening the conditions for marginalized communities. By pursuing data science intelligence, we are able to correct for these failures. The participation, access, inclusion and representation (PAIR) principles attempt to mitigate these negative effects by embedding diversity and equity into the data science discourse.

References

Arend, M. 2019. "Business Climate Rankings: VII Straight," Site Selection Magazine, November, 2019 (https://siteselection.com/issues/2019/nov/business-climate-rankings-seven-straight-georgia-sets-a-record.cfm; accessed: November 9, 2019)

Azam, A. 2014. "The First Rule of Data Science," *Berkeley Science Review*. (http://berkeleysciencereview.com/article/first-rule-data-science/; accessed: November 5, 2019)

Bayern, M. 2019. "Why Only 18% of Data Scientists Are Women," TechRepublic, May 20, 2019 (https://www.techrepublic.com/article/why-only-18-of-data-scientists-are-women/)

Berman, F., Rutenbar, R., Hailpern, B., Christensen, H., Davidson, S., Estrin, D., Franklin, M., Martonosi, M., Raghavan, P., Stodden, S., and Szalay, A. S. 2018. "Realizing the Potential of Data Science," *Communications of the ACM*, (61:4), pp. 67-72 (doi: 10.1145/3188721).

Boykis, V. 2019. "Data Science is Different," (http://veekaybee.github.io/2019/02/13/data-science-is-different/)

Bozeman, B. 2002. "Public-Value Failure: When Efficient Markets May Not Do," *Public Administration Review*, (62:2), pp. 145-161.

Bozeman, B. 2003. "Public Value Mapping of Science Outcomes: Theory and Method," Knowledge Flows and Knowledge Collectives: Understanding the Role of Science and Technology Policies in Development, (2), pp. 3-48.

Bozeman, B. (2007). *Public Values and Public Interest: Counterbalancing Economic Individualism*. Washington, DC: Georgetown University Press.

Bozeman, B. and Johnson, J., 2015. The political economy of public values: A case for the public sphere and progressive opportunity. *The American review of public administration*, (45:1), pp.61-85 (https://doi.org/10.1177/0275074014532826)

Buolamwini, J., and Gebru, T. 2018. "Gender shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Conference on Fairness, Accountability, and Transparency*, (81) pp. 77-91.

Bush, V. 1945. *"Science, The Endless Frontier: A Report to the President,"* US Government.

Cath, C. 2018. "Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges," *The Royal Society Publishing*, (doi: 10.1098/rsta.2018.0080).

Chui, M., Bughin, J., Hazan, E., Ramaswamy, S., Allas, T., Dahlström, P., Henke, N., Trench, M. 2017. "*Artificial Intelligence the Next Digital Frontier?*," McKinsey and Company Global Institute, (47).

Cozzens, S. E. 2007. "Distributive Justice in Science and Technology Policy," *Science and Public Policy*, (34:2), pp. 85-94.

Croll, A. 2012. "Big Data is Our Generation's Civil Rights Issue, and We Don't Know it," *Big data now*, pp. 55-59.

Daniels, J., Nkonde, M., and Mir, D. 2019. "Advancing Racial Literacy in Tech," Data & Society's Fellowship Program (https://racialliteracy.tech/; accessed: November 5, 2019)

Deng, X., Joshi, K. D., and Galliers, R. D. 2016. "The Duality of Empowerment and Marginalization in Microtask Crowdsourcing: Giving Voice to The Less Powerful Through Value Sensitive Design," *MIS Quarterly*, (40:2), pp. 279-302.

Dietterich, T. G. 2017. "Steps Toward Robust Artificial Intelligence," *AI Magazine*, (38:3), pp. 3-24 (doi:10.1609/aimag.v38i3.2756).

Duff, A. S. 2005. "Social Engineering in The Information Age," *The Information Society*, (21:1), pp. 67-71 (doi: 10.1080/01972240590895937).

Ferrara, E., Varol, O., Davis, C., Menczer, F. and Flammini, A. 2016. "The rise of social bots," *Communications of the ACM*, (59:7), pp. 96-104 (doi:10.1145/2818717)

Finzer, W. 2013. "The Data Science Education Dilemma," *Technology Innovations in Statistics Education*, (7:2).

General Assembly. 2017. "The Study of Data Science Lags in Gender and Racial Representation," *The Index,* September 25, 2017 (https://theindex.generalassemb.ly/data-science-education-lags-behind-in-diversity-ff59ffa718ec; accessed: September 12, 2019)

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, (78: 6), pp. 1360-1380.

Harnham Report. 2019. "USA Diversity in Data and Analytics: A review of diversity within the data and analytics industry in 2019," (https://www.harnham.com/us/2019-usa-diversity-in-data-analytics-report; accessed: September 12, 2019)

Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., and Sethupathy, G. 2016. "The Age of Analytics: Competing in a Data-Driven World," *McKinsey Global Institute*, pp. 1-136.

Hill Happenings. March 2019. Dorothy Vaughn Tech Symposium. (https://www.hillhappenings.com/list/2019/3/6/dorothy-vaughan-tech-symposium)

Kleinberg, J., Ludwig, J. and Mullainathan, S. 2016." A guide to solving social problems with machine learning." *Harvard Business Review*. (https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning)

Kneese, T., Rosenblat, A., and Boyd, D. 2014. "Inequalities and Asymmetries the Social, Cultural & Ethical Dimensions of "Big Data," (http://dx.doi.org/10.2139/ssrn.2538558)

Kross, S., Peng, R. D., Caffo, B. S., Gooding, I., and Leek, J. T. 2017. "The Democratization of Data Science Education," *PeerJ Preprints*.

Leslie, D., and Richard, F. (2019). "College Faculty Have Become More Racially and Ethnically Diverse, but Remain Far Less So Than Students," *Pew Research Center*. (https://www.pewresearch.org/fact-tank/2019/07/31/us-college-faculty-student-diversity/; accessed November 9, 2019)

MacPhail, T. 2015. "Data, Data Everywhere," *Public Culture*, (27:2 (76)), pp. 213-219.

Mantha, Y., and Hudson, S. 2018. "Estimating the Gender Ratio of AI Researchers Around the World," Medium's Element AI Blog, April 18, 2018 (https://medium.com/element-ai-research-lab/estimating-the-gender-ratio-of-ai-researchers-around-the-world-81d2b8dbe9c3; accessed: November 5, 2019)

Marshall, B. 2017. "Data Science Experiences for Undergraduates," *Journal of Computing Sciences in Colleges*, *33*(2), pp. 198-204.

Marshall B., and Geier S., 2019 "Targeted Curricular Innovations in Data Science," *Proceedings of the IEEE Frontiers in Education Conference*.

McNie, E. C., Parris, A., and Sarewitz, D. 2016. "Improving the Public Value of Science: A Typology to Inform Discussion, Design and Implementation of Research," *Research Policy*, (45:4), pp. 884-895.

National Academies of Sciences, Engineering, and Medicine. 2018. "*Envisioning the Data Science Discipline: The Undergraduate Perspective,*" Washington, DC: The National Academies Press. (doi:10.17226/24886).

Noren, L., Helfrich, G., and Yao, S. 2019. "Who's Building Your AI? Research Brief," Obsidian Blog, October 30, 2019 (https://www.obsidiansecurity.com/whos-building-your-ai-research-brief/; accessed: November 5, 2019)

North, D. C. (1991). "Institutions," *The Journal of Economic Perspectives, 5*, pp. 97-112.

O'neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York, NY: Crown Publishing Group

Price Waterhouse Coopers. 2017."Investing in America's Data Science and Analytics Talent: A Case for Action," *Business-Higher Education Forum Report*.

Schwab, K. 2016. "The Fourth Industrial Revolution," *World Economic Forum*. Geneva, Switzerland.

Song, I. Y., and Zhu, Y. 2016. "Big Data and Data Science: What Should We Teach?," *Expert Systems*, (33:4), pp. 364-373.

U.S. Census Bureau. 2018. State of Georgia. (https://www.census.gov/quickfacts; accessed: November 6, 2019).

U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System (IPEDS). 2016. State of Georgia (https://nces.ed.gov/ipeds/use-the-data; accessed: November 5, 2019).

Uzzi, B. (1997). Social Structure and Competition in Interfirm Networks. *Administrative Science Quarterly*, (42:1), pp. 35-67.

Vosoughi, S., Roy, D. and Aral, S. 2018. "The spread of true and false news online," *Science*, (359:6380), pp. 1146-1151.

Wellman, B., 2001. Computer networks as social networks. *Science*, (293: 5537), pp. 2031-2034.

Wolf, C., 1979. "A theory of nonmarket failure: framework for implementation analysis." *Journal of law and economics*, pp. 107-139.