Artificial Intelligence and Social Responsibility: The Roles of the University

A white paper by faculty members at the *University of Illinois Urbana-Champaign*
Nigel Bosch, *School of Information Sciences and Department of Educational Psychology*
Anita Say Chan, *School of Information Sciences and College of Media*
Jenny L. Davis, *American Indian Studies and Anthropology*
Rochelle Gutiérrez, *Department of Curriculum and Instruction*
Jingrui He, *School of Information Sciences*
Karrie Karahalios, *Department of Computer Science*
Sanmi Koyejo, *Department of Computer Science*
Michael C. Loui, *Department of Electrical and Computer Engineering*
Ruby Mendenhall, *Sociology, African American Studies, and Carle Illinois College of Medicine*
Madelyn Rose Sanfilippo, *School of Information Sciences*
Hanghang Tong, *Department of Computer Science*
Lav R. Varshney, *Department of Electrical and Computer Engineering*
Yang Wang, *School of Information Sciences*

November 7, 2022

Summary

Contemporary AI technologies are more powerful and pervasive than the original AI technologies created in university laboratories. While industry dominates AI today, universities can still play important roles. In this white paper, we recommend actions that universities should take to promote social responsibility in the development and application of AI technologies. Our paper addresses four key questions connected with the university missions of education, research, community engagement, and public service:

1. How can universities effectively educate technical professionals and the public to consider social responsibilities in the design and use of AI systems?
2. How can university and industry researchers collaborate on AI technologies in a socially responsible way?
3. How can universities better collaborate with external organizations and local communities to address questions of bias and discrimination in AI technologies?
4. How can universities contribute to the governance of AI technologies?

Overall, we call for greater engagement between universities and external stakeholders, in which academics collaborate with industry practitioners, government policymakers, and community partners. These collaborations can promote social responsibility by ensuring that AI technologies are responsive to community needs, rather than driven solely by business interests. Universities should also ensure that in the criteria for promotion and tenure of faculty members,

teaching interdisciplinary courses and building external networks are recognized as valuable forms of scholarship.

This white paper reflects ideas that were presented and discussed on March 23, 2022 at the Symposium on Artificial Intelligence and Social Responsibility sponsored by the Coordinated Science Laboratory at the University of Illinois Urbana-Champaign, with co-sponsorship by the School of Information Sciences at the University. This symposium was the second of two symposia sponsored by the Laboratory to celebrate its 70th anniversary. The earlier symposium produced a companion white paper on the future of computing, which appeared in the Computing Community Consortium Blog on September 27, 2022.

Introduction

Developments in artificial intelligence (AI) are affecting almost every corner of our society at an unprecedented speed and scale. While the economic and social benefits of AI technologies have been celebrated, many harms of AI have also been documented, such as racial biases in recidivism models used for criminal sentencing, and gender discrimination in automated screening of job applications (Banaji et al. 2019; Benjamin 2019b; Couldry and Mejias 2020; Crawford 2021; Eubanks 2019; O'Neill 2016; United Nations Human Rights Council 2018; Zuboff 2019). Some AI technologies use massive collections of personal data, which can lead to violations of privacy and losses of intellectual property (Kang et al. 2021). In numerous cases, AI introduces or amplifies biases that disproportionately impact historically marginalized and vulnerable communities, including ethnic and religious minorities, LGBTQ populations, and individuals who are Black, Indigenous, and/or people of color (Banaji et al. 2019; Benjamin 2019a; Buolamwini and Gebru 2018; Chun 2021; Noble 2018; United Nations Human Rights Council 2018). As AI changes our world, its benefits may not be equally shared among future generations (D'Ignazio & Klein 2020; Kandlhofer et al. 2016; Hermann 2021).

AI technologies today are far more powerful than the early AI technologies created in universities. We ask: what roles should universities now play, given that the stakes, both the benefits and the harms, of AI are extremely high? While many organizations have published statements of principles for the socially responsible development of AI technologies (Saveliev and Zhurenkov 2020; Krkač and Bračević 2020; Darnault, Parcollet, & Morchid 2019; Santoni de Sio & Macacci 2021; Cath et al. 2018), in this paper, we move beyond statements of principles to recommendations for actions by universities. We focus on four key questions connected with the university missions of education, research, community engagement, and public service:

1. How can universities effectively educate technical professionals and the public to consider social responsibilities in the design and use of AI systems?
2. How can university and industry researchers collaborate on AI technologies in a socially responsible way?
3. How can universities better collaborate with external organizations and local communities to address questions of bias and discrimination in AI technologies?
4. How can universities contribute to the governance of AI technologies?

To explore these issues, the Coordinated Science Laboratory at the University of Illinois Urbana-Champaign hosted a Symposium on Artificial Intelligence and Social Responsibility on

March 23, 2022. The symposium was co-sponsored by the School of Information Sciences at the University. The symposium planning committee consisted of six of the authors of this white paper: Chan, Koyejo, Loui, Tong, Varshney, and Wang. Authors Bosch, Davis, Gutierrez, He, Karahalios, Mendenhall, and Sanfilippo served as respondents to the featured lectures by Mary Gray, a senior principal researcher at Microsoft Research; Kush Varshney, a researcher and manager at IBM Research AI; David Kaiser, a professor of the history of science and professor of physics at the Massachusetts Institute of Technology; and Suresh Venkatasubramanian, a professor in computer science and data science at Brown University. Video recordings of the lectures and discussions are available at the symposium website (https://csl.illinois.edu/events/ai-social-symposium).

Education

Education is a key mission of universities, ranging from formal classroom instruction to engagement with practitioners, policymakers, and publics beyond the campus. The educational imperative is especially critical for AI and social responsibility because the real-world stakes are high: although education broadly benefits society, there is a substantial risk of harm to individuals and communities if education leads to more widespread application of AI (e.g., Baker and Hawn 2021; Kirkpatrick 2016). A student learning about AI today can apply computational power and algorithmic methods with ease unheard of even a decade ago, and effect enormous impacts on society through data-driven products and analyses—for example, by applying cloud computing resources to train large natural-language processing models that serve as chatbots or product review evaluators in e-commerce applications, or even by pursuing intentionally harmful purposes, like creating deceptive news content with fake images using DALL-E, at a massive scale.

Who should be involved?

Today's students are building our future. Many already possess concerns about a troubled world that they are inheriting (e.g., Black Lives Matter movement, LGBTQ rights, Indigenous land rights, climate change, COVID-19, and a divided public) and they seek to make real changes in society. However, we cannot assume that members of various communities will learn "on the job," or from life, to consider the societal impacts of AI technologies and reimagine possible futures; they need structured opportunities to develop skills and sensibilities now so that they can make informed choices about where and how to work responsibly, challenge unethical practices that arise, and help lead society.

Given the broad population that can benefit from such educational efforts, we note that these efforts should support various types of educators and learners, such as K–12 students, college students, technical professionals who seek continuing education credits, lifelong and self-directed learners (e.g., homeschooled children or hobbyists), and the general public. The instructional materials should not be designed to focus only on future computer scientists and technical professionals, but rather be tailored to students from different fields of study across the curriculum.

In addition, education researchers need to take up issues regarding social responsibility in AI education. These issues include, for example, what is known and what needs to be studied, what constitutes "literacy" in AI (i.e., learning outcomes), and what is effective for whom (not just

for engineering students learning AI techniques, but also for students who will be affected by AI). As much as AI and society are constantly evolving, so do these needs. If the goal is social transformation, we should assume those in power will push back. Educators and future workers will thus need to be prepared with the skills and sensibilities to articulate their values to others (Gentile 2010), know how to navigate politics in work environments, and understand how to carry out this work collectively. Having opportunities to "try out" skills, such as via role play (Loui 2009), will be invaluable.

How do we do it?

In colleges and universities, issues in AI and social responsibility are currently covered in course modules (Grosz et al. 2019), in courses on AI and ethics (Burton et al. 2017), and in courses on computing ethics (Fiesler, Garrett & Beard 2020). Cambridge University offers a course leading to a master's degree in AI Ethics and Society (Institute of Continuing Education 2022). In these courses, topics and teaching methods have included ethics codes, philosophical theories, and discussions of historical events and science fiction. These courses are sometimes taught by multidisciplinary teams, with members from computing, data sciences, humanities, arts, and social sciences. One example of multidisciplinary instructional collaboration is the Social and Ethical Responsibilities of Computing (SERC) initiative at the Massachusetts Institute of Technology (MIT 2022); case studies developed by SERC are freely available online. To date, however, there has been little formal, empirical research on computing ethics education. Reviewing the limited research, Hedayati-Mehdiabadi (2022) called for more studies on the effectiveness of different pedagogical approaches. These studies would need valid, reliable assessment instruments to measure what students have learned, but no assessment instruments have yet been developed for AI ethics (Goldsmith et al. 2020).

We recommend that universities extend these instructional efforts to reach a wide range of audiences, including non-specialists, in settings beyond schools, such as lifelong-learning programs, local libraries, museums, places of worship, and other community spaces. Like the SERC initiative, instructional materials should be freely available online (e.g., with Creative Commons licenses) and inclusive (e.g., to people with disabilities). To reach diverse audiences, these materials should have multiple modalities (e.g., blog posts, podcasts, and short videos) and include hands-on activities (e.g., co-writing, role-playing, and design activities). These materials can be shared via sites such as the Online Ethics Center for Engineering and Science at the University of Virginia (Online Ethics Center 2022). This Center not only promotes the sharing, adoption, and adaptation of peer-reviewed instructional materials, but also supports communities of practice in which instructors can share success stories, reflect on the influence of local contexts, and discuss dilemmas of practice.

We recognize that multidisciplinary instructional collaborations are not always valued by university reward structures. Therefore, it is helpful for these instructional teams to include members from a range of ranks and positions: senior faculty members can signal the importance of such efforts and can help align these efforts with the career goals of younger colleagues, to ensure that younger scholars will receive credit toward promotion.

Research

University and industry researchers collaborate in various ways, including the sharing of industry data and the employment of doctoral students as industry interns. Such collaborations require advance agreements around intellectual property, data sharing, scheduling, conflicts of interest, and more. These institutional and researcher alignments allow for common ground and the creation of research workflows. However, there are ethical differences between industry and academia with respect to research oversight, data repositories, data access, and research integrity.

In the following, we begin by discussing ethical infrastructures (e.g., the Common Rule) that guide research at federally funded institutions such as universities (and a few companies) and how they address data collection, consent, privacy, and access with respect to AI systems. We then discuss responsible AI systems through the lens of AI explainability. Finally, accountability practices for both the data used in AI systems and explainability help signal the reasoning and values embedded in an AI system and in the research process.

Since research projects in social computing and computational social science have resulted in numerous unintended consequences, a modified ethical research framework may be one path forward to align research practice and values across the academy and industry. More work needs to be done to address data governance, AI transparency, and equitable data access.

Challenges to Institutional Review Board (IRB) adherence around data, use, and inference

Institutional Review Boards (IRBs) began in 1974 with the signing of the National Research Act, with the goal of protecting human subjects in biomedical and behavioral research (Moon 2009). The National Research Act was a response to the ongoing problem of unethical research projects despite the presence of clear standards for ethical research, such as the Nuremberg Code. This led to the creation of the *Belmont Report* (National Commission for the Protection of Human Subjects 1979) and its three foundational principles:

1. *Respect for persons*. Individuals should be treated as autonomous agents with a right to privacy, and vulnerable persons should be protected.
2. *Beneficence*. Persons should be treated ethically, respected, and protected from harm. Moreover, possible harms should be minimized, while maximizing possible benefits.
3. *Justice*. The benefits and burdens of research should be shared equally across involved persons.

Most universities today have IRBs or equivalent boards, but most industry firms do not, with a notable exception being the Ethics Review Program at Microsoft Research. Without human-centered data, formal monitoring of research involving human subjects is not required. For example, creating a dataset of identifiable plant leaves for a plant recognition AI algorithm does not require IRB approval. However, creating a dataset of identifiable people would require IRB approval, as they could be identified (“respect for persons”), such identification might cause them harm (“beneficence”), and they might be more prone to harm than others (“justice”).

Data are at the core of machine learning (ML) and AI models. The reason is that ML models (such as those using neural networks) use a primary set of data, known as a “training set,” to “learn,” and these training sets help to create a model that can then be used with other data as

input. When data involve human subjects, protections such as those discussed by IRBs are critical, even when the data collection activities do not fit strict definitions of "research." Yet many challenges around data collection and use remain, from lack of IRB oversight to unanticipated consequences. Precedent suggests that we will continue to encounter unintended consequences even with IRB oversight.

In the past, large collections of data have caused violations of privacy. For example, a 2006 AOL dataset containing the browser search queries of individuals unintentionally revealed their identities (Zimmer 2020). Because private identifiable information was inferred, this practice violated the Federal Policy for the Protection of Human Subjects, known as the Common Rule.

In another example, after the Cambridge Analytica firm collected data harvested from over 87 million Facebook users, the dataset was used to target potential voters in 2016 political campaigns. However, the data (again, used for political purposes) were collected by a university-affiliated data scientist via a game-like personality quiz app called "This is Your Digital Life." That app inferred personality profiles of its users *and* collected additional third-party information about their Facebook connections. Users participated by answering a series of questions, but did not consent to offer their data to other entities (or used for commercial purposes), nor did their online friends consent to the collection of their data. Cambridge Analytica's collection of the data via Facebook violated people's privacy, employed unapproved recruitment practices, and involved unapproved data transfer under a university's oversight. Moreover, the case raised issues of how IRBs should work when university-affiliated entrepreneurs are involved. After initially arguing that people willingly offered their information, Facebook suspended the app a few years later and was issued a landmark $5 billion fine for privacy violations and mishandling of data (Ma and Gilbert 2019).

Many similar cases have occurred and will likely continue to emerge as new online collection practices appear faster than IRB and external guidelines and laws can be created to regulate them. The concerns around privacy, procedural ethical adherence, deception, and reasonable expectations are directly addressed in the Belmont Report and the Common Rule via their three pillars: integrity, beneficence, and justice.

Unintended consequences will likely continue to happen. It recently emerged, for example, that the company Clearview has collected over 20 billion images of people online without consent. These images are being sold for purposes known and unknown (Clearview 2021), and we know we won't discover them immediately. How can we move forward when we can't fully predict consequences and want to protect people and communities? Given the pervasiveness of ethical challenges, we should address the benefits and harms to society, in addition to individuals, as expressed in the three Belmont principles. More and more, we are faced with data-centered challenges that affect society as a whole. Given the delay between studies/deployments and unintended consequences, this framing results in questions for both industry and academia. First, how do we define privacy protection given that deanonymization is frequently possible (as discussed in the earlier AOL example) (Zimmer 2020)? Second, how should we move forward to protect privacy, given the challenges?

Datasets: bias and governance

Much recent work cites dataset challenges in AI technologies that employ ML. Questions around bias and equity abound. Crawford (2017) highlighted allocative and representational harm in datasets with biased collection practices. "Allocative harm" refers to harm resulting in inequitable allocation of resources or opportunities to one group: for example, offering loans at better rates to one group over another. "Representational harm" results in discrimination, as when a Google search for the term "gorilla" returned images of Black people. Others point out incongruences when data are collected for one purpose and used for another (as in the Cambridge Analytica case), and argue that we must understand data provenance: who collected the data, when, why, and for what purpose (D'Ignazio & Klein 2020)?

Further, when datasets result in biased outcomes, what is to be done? How should we govern diverse datasets, to decide who can access them, whether they should be accessed, who can augment them, and so forth? Let's consider the case of the ImageNet dataset (ImageNet 2021). The dataset, created by researchers at Princeton and Stanford, contains over 14 million images labeled by over 30,000 workers on Amazon's Mechanical Turk platform. Freely available on a website for noncommercial use, this dataset has enabled controlled studies that compared different ML algorithms and has been used in over 300 research papers. However, researchers have discovered various forms of bias in this dataset. One example is that greater pleasantness of an image is associated with lighter skin tone (Wiggers 2020). Reasons likely include unbalanced data collection, biased labeling, and more. Efforts are underway to better understand the biases within the dataset, and questions remain. Given the known biases, how should this dataset be presented? Should it be available? How will people who downloaded it a decade ago find out that it is biased? How should versions of this dataset be labeled or archived?

While concerns remain around datasets collected without scientific rigor, sometimes research experts and corporations come together to curate a quality dataset, as was the case with Social Science One (2022), a consortium created as a model for academic-industry collaboration. A proposal-based model gave researchers with approved proposals access to Facebook data pertinent to elections and democracy. However, there was an error in the data collection: data from U.S. users with no political leaning, who represented roughly *half* of Facebook users, were accidentally omitted. This omission affected the ongoing work of at least 110 researchers (Timberg 2021), the dozens of papers published based on the data, and trust around datasets that cannot be verified by researchers.

Managing the complexity of such datasets—organizing them, archiving them, and validating them—requires new methods and regulations. Historically, libraries collected, archived, and cataloged data collections (e.g., newspapers, books, and music). Now, in addition to libraries, corporations and start-ups collect such data, and maintaining these data for public consumption requires ongoing funding. Addressing such ongoing data needs requires the collaboration of academics, industry, and policy experts. We should address issues of data versioning, data integrity, data ethics, data access, and data verification. Datasets such as those curated by Social Science One have the potential to help us assess our democracy. These datasets should be audited, and systems using them should provide explanations and reasoning. That will require insights from academics who use the data and from industry experts who provide it.

Explanations/reasoning and social impacts (Littman et al. 2021)

The opaque nature of many existing AI and ML systems can cause distrust and raise privacy concerns among users, because users of AI and ML systems often do not understand how these systems reach conclusions and make decisions. Unlike other technologies whose mechanisms are hidden but that have been tested over a long period of time (e.g., car engines), AI and ML's use in many high-stakes application domains (e.g., healthcare and criminal justice) is relatively new, and in multiple incidents, AI and ML systems have produced great social harms. For example, in 2016, it was discovered that an AI tool used in courtrooms across the U.S. to predict future crimes, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), was biased against Black defendants. This led to widespread criticism and concerns among the general public. Therefore, from both research and practical perspectives, it is critical to address this problem.

To address the opaque nature of many existing AI and ML systems, government agencies have invested in a variety of high-profile programs intended to promote model explainability and interpretability, such as the DARPA Explainable Artificial Intelligence Program and the NSF Fairness in Artificial Intelligence Program. Because of the rapidly growing societal impacts of AI and ML systems across multiple high-impact application domains, close and effective collaboration is needed between academia and industry to address the opaque nature of these systems. Recent years have seen more creative forms of such collaboration, ranging from academic research centers sponsored by industrial partners (e.g., the IBM-Illinois Discovery Accelerator Institute) to industry-sponsored academic research projects.

In the past, university researchers have mostly pursued fundamental research without immediate applications. With academia-industry collaborations, however, university researchers can now access large volumes of user data collected by industry. This is quickly changing expectations of what research can be done by academia, as more university researchers are involved in the design, application, and deployment of AI and ML systems. One of the traditional patterns of academia-industry collaboration has been for industry to identify challenging research problems present in their business operations, and for academic researchers to join the effort to solve these problems. However, that pattern might limit the range of problems being studied through such collaborations, especially those that involve potential conflicts of commercial interests. For example, the best way of explaining and interpreting system outputs from the users' perspective may not be acceptable from the company's perspective, because of a potential leak of proprietary information. Therefore, it is critical to identify other collaboration patterns with potential involvement of third parties, to address societal issues arising from the deployment and application of AI and ML systems.

In addition, to facilitate effective and efficient model explanation and interpretation, it is highly recommended that various educational components targeting different populations be seamlessly integrated into both academic programs and company products. These educational components across academia and industry should ensure that the users are aware of their rights when certain services are rendered, and that the explanation and interpretation of AI and ML systems meet the constantly evolving user needs. For example, through the educational components, users would gain basic knowledge of what is meant by "one feature plays a critical role in the decision by an AI system," and of whether their privacy has been compromised if the

decision depends on certain attributes of their profiles. In particular, more efforts are needed for the educational components to reach marginalized populations, because of the limitations of existing outreach and broadening-participation activities.

Actionable areas for collaboration

We've discussed several challenges around the collection, protection, and use of data that pertain to people. We briefly summarize some specific areas in which academics and industry experts can jointly move forward.

(1) Policies for data collection: Working with social scientists, statisticians, and data scientists to incorporate representative data collection methods and make clear for what purpose data were collected.
(2) Practices for archiving/access: Understanding existing archival techniques and how they can evolve to capture the nuances of our evolving data landscape.
(3) Algorithm and data audits: Techniques to validate datasets and algorithms such that they behave in the expected manner. (We must explore who should conduct these audits, when, and how.)
(4) Explanation: Techniques to reason with systems to better understand why they do what they do, so we can better understand the reasoning behind decisions.
(5) Understand conflicts of interest: While industry-university collaboration can help us move forward, it is important to note that there have been, and continue to be, challenges around access. Can we provide equitable data access to researchers, and how? (Note also that some researchers choose not to collaborate with entities with questionable ethical practices.) And when collaborations do occur, how can we mitigate conflicts of interest? Should a company pay an academic researcher for an audit of their system? Would that researcher return a favorable audit report if the company funds their research?
(6) Processes to address social impacts: To study and anticipate future consequences to create mitigation strategies. Conferences have begun requiring speculation on possible future use outcomes that may result from ML research. This has been met with both acceptance and reluctance, and we know more work needs to be done. One avenue for exploration is speculative design (Dunne and Raby 2013), which has shown promise in several spaces (Eslami et al. 2018; Fiesler 2019).

While concerns around matters such as intellectual property and conflicts of interest will not disappear, for academia and industry to communicate effectively and come together to create new collaboration models such as Social Science One (2022), we must first share a common set of principles and values. Given today's societal challenges of AI technologies that affect our health and our democracy, a common framework adopted by both academia and industry would create a common ground around research practice.

Community Collaborations

Tech industry actors, granting agencies, and academic institutions have frequently conceptualized AI technologies as universal solutions to an ever-broadening array of social problems. Such conceptualizations have been critiqued for drawing on and perpetuating colonial

orientations (which include political and economic exploitation) by positioning AI as the horizon of an evolutionary arc towards an inevitable future of progress and enlightenment and away from a "primitive past" that is often associated with racialized and marginalized communities and groups (Couldry & Mejias 2019). These overarching discussions imagine technological innovation as the outcome of individual (usually white and male) genius in the urban, middle-to-upper-class global north that is then "brought to" marginalized communities (Chan 2014). Mainstream technology research and development practices further limit access to resources and problem-solving capacity to research populations largely privileged along lines of geography, gender/sexuality, race, class, and ability, among other social characteristics, leading to research problems over-defined by white middle-class males in the urban North. (Brown et al. 2016; Brown et al. 2019). These AI practices perpetuate systemic exclusions.

The current AI ecology narrowly represents existing social interests and concerns around technology (Benjamin 2019b; Broussard 2018). This AI ecology and culture often creates systems that lead to repeated over-policing and over-surveillance of marginalized (i.e., socially and economically excluded) individuals and groups (Benjamin 2019a; Browne 2015; Jefferson 2020; O'Neill 2016). In addition, researchers participating in AI ecology and culture routinely exploit and extract data about the experiences of historically marginalized and underrepresented groups, such as Black, Indigenous, Latinx, Asian, and Middle Eastern populations (Cifor et al. 2019). Left unaddressed, this practice of simultaneous extraction and exclusion represents a crisis in institutional knowledge production. This crisis can only be addressed when researchers transform their practices and accept accountability for past and ongoing harms. It is critical for researchers to adopt methodologies of collaboration, care, and reparations when working with marginalized communities. Researchers must recognize that all data and research are produced in political and historical contexts that are shaped by systems of cultural beliefs (Lewis et al. 2018).

Local communities and external organizations that are led by and represent the interests of historically marginalized and vulnerable groups are critical to the work of reimagining AI research in ways that fundamentally center methods of care and reparation and address questions of bias and discrimination. That is the case not only because marginalized populations have been disproportionately harmed by AI technologies (Buolamwini & Gebru 2018; Chun 2021; Eubanks 2019; Noble 2018; McIlwain 2020; O'Neill 2016), but because they have always been the sources of alternative models of innovation, knowledge practice, and future-making (Benjamin 2019a; Brock 2020; Chan 2021; Costanza-Chock 2020; Davis 2017). Local communities and grassroots networks have long been up to the work of innovation (Eubanks 2011), whether they are developing mutual aid networks—such as Black, Indigenous, immigrant, feminist, and LGBTQ alternative health care networks, or alternative research and education initiatives developed by disability rights advocates to support underrepresented communities—to hack and improvise workarounds for existing technological systems that were not designed to meet their needs (D'Ignazio and Klein 2020; Mendenhall et al. 2017b; Nelson 2013), or figuring out how to scale existing infrastructural resources to meet the under-met demands of nurturing and care in ways that nurses, public school teachers, and other care workers do daily (Precarity Lab 2020; Varshney 2022). Much of this work has been ignored or undervalued by "innovation" frameworks that celebrate commercial, profit-generating high-tech products (Broussard 2018; Crawford 2021) and that problematically frame local communities as mere sources of problems to be solved,

or relevant primarily as sources of experimentation or objects of research. Marginalized populations, however, have themselves led the innovation of new knowledge practices centered on the needs, interests, and concerns of the people most directly harmed by dominant norms around knowledge production (Chan & Garcia, forthcoming). Whether in the late 19th-century feminist and immigrant-authored surveys and labor studies of Hull House (Chan 2020), in the early 20th-century data journalism of Ida B. Wells, in the data visualizations of W.E.B. DuBois (Battle-Baptiste & Rusert 2018; Mendenhall et al. 2017a), or in the mid-20th-century origins of accessibility design and educational research at the University of Illinois Urbana-Champaign (Brown 2008, Reagan 2017), marginalized communities have worked to create new research futures that center care, restoration, resilience, repair, and reparation (rather than innovation or economic growth) as their primary objectives and outputs.

Contemporary community organizations—including the many that routinely undertake research and data collection in their work for local populations—continue to follow in those footsteps. Community members and organizations, nonprofits, and citizen groups, in short, offer a wealth of knowledge and expertise grounded in empirical conditions, and can be included and empowered throughout the research process in the work to build AI futures around the principles of justice. These groups—spanning such collaborative projects as the Our Data Bodies initiative, Data for Black Lives, and the Global Indigenous Data Alliance—have actively built multi-sectoral collaborations to undertake research that tracks the negative impacts of AI technologies in marginalized communities, and to challenge and mitigate algorithmic bias and discrimination in a variety of locales (Brown et al. 2019; Carroll et al. 2020; Chan 2021; Irani 2021; Lewis et al. 2018; Milner 2020; Petty 2018). Universities can play an important role in collaborating with external organizations and communities to support and amplify such work (Davis 2022; Flores et al. 2014; Ginsburg 2019), but these collaborations will require genuine transformations of institutional norms, structures, and standard academic practices in order to effectively tackle questions of bias and discrimination in AI, and to cultivate approaches to technological infrastructures from community-centered perspectives.

It will require support for processes—and a recognition and valuation of the added labor needed from campus actors and faculty in particular—to build genuinely reciprocal partnerships with community members in ways that disrupt conventional valuations and hierarchies between campus scholars and off-campus civic researchers as knowledge practitioners. It will require support too for processes that redefine and transform disciplinary norms that have traditionally encouraged scholars to conduct research independently, that imagine research engagements as primarily developed within a single discipline for disciplinary audiences, and that see research as primarily led by professional scholars. It will also require patience from community and campus actors alike as new work is invested to build trust, explore and define mutually supportive research benefits, and potentially recognize past harms from previous research undertaken in developing AI technologies. The institutional transformations required for campuses will undoubtedly take time and sustained support from leadership at every level. As initial steps, university leaders should recognize publicly engaged scholarship in evaluation and promotion criteria for faculty and staff, and leaders should provide opportunities for campus colleagues already engaged in community-empowering research collaborations to collectively organize strategies for growing data partnerships and AI developments towards healing and emancipatory

futures. Community-empowering partnerships will open new challenges for universities and knowledge institutions, but they will also open extraordinary opportunities for campus and civic scholars to explore new research possibilities that stretch beyond singular worldviews, ways of knowing, and forms of accountable knowledge production to press towards a future of shared dignity, well-being, and justice.

Governance

Introduction

While advances in AI technologies have accelerated, the development of regulations and ethical guidelines has lagged—most conspicuously in the United States (Buiten 2019; Khisamova, Begishev & Gaifutdinov 2019; Yara et al. 2021; Wu 2018). Present AI governance consists primarily of fragmented regulatory interventions that capitalize on reactions to industry failures and harms to individuals, as well as ethical principles (Dafoe 2018), and reflect the industry-specific interests and preferences of the most powerful actors (Jung and Sanfilippo 2022). As technological innovation accelerates, regulatory change is slow, and ethical guidelines, when placed in tension with profit margins, are often ineffective. AI governance has been most impactful when coordination occurs between stakeholders (Varshney, Keskar & Socher 2019), as with the Partnership on AI (2022), and across systems or domains, as with contextually flexible frameworks like the draft NIST framework on AI Risk Management (2022). Governance is more than regulation, management, or standards (e.g., C/S2ESC - Software & Systems Engineering Standards Committee 2021); governance also encompasses social norms, markets, and choice architectures (Lessig 1999; Thaler, Sunstein & Balz 2013), among other forms of institutional structure (Williamson 1996). Society needs more coordination and responsiveness to community and social needs, rather than data- or profit-driven responses alone, to promote social good and prevent harms. Sustainable governance for AI will co-evolve with technological change.

Academia can and should fill this role by encouraging interdisciplinary scholarship, recognizing the labor required to connect stakeholders, incentivizing translation of research into practice and public outreach, and fostering financial independence and intellectual freedom to enable academics to provide critiques and serve as outside change agents. Universities are well situated for this role, given their centrally networked location among policymakers, government, courts, people and communities, media, and industry. However, they must make a good-faith effort to be both independent and trustworthy in their engagement with marginalized communities. Here, we outline how universities can strategically respond to and propagate social norms, as well as contribute to responsive rule-making, by sharing expertise, amplifying voices, auditing systems, and advising on best practices (Venkatasubramanian, 2022).

Sharing expertise

Research universities have a mission to address complex, pressing questions from theoretical and empirical perspectives across basic science and practical applications, and to educate broadly (Madison 2020). This mission, which is especially urgent with respect to rapid AI innovation, ideally transcends the incentive structure of the private sector and is independent of the political interests of the public sector. The result is a wealth of expertise developed via

research, experience, and networked relationships across sectors. Universities as institutions and individual academics have ties to industry, government, civil society, and communities via their alumni network, research collaborations, and partnerships on various initiatives. Academics lend their expertise through expert testimony, public comment periods, and open scholarship, while also cultivating relationships with experts outside the academy. Universities are key to brokering and facilitating engagement among experts and stakeholders in discussion of AI governance, as well as physical spaces to host those discussions. Universities should open more events and knowledge resources to the public; democratize access and information; leverage their networks when called upon to address pressing challenges (e.g., COVID-19 responses); recognize the validity and importance of these activities in promotion and tenure; and instill the logic of representative and participatory decision-making in their students, as a key principle of general education, to prepare them to participate actively in co-evolution of innovation and governance.

Amplify voices

Academic researchers often recognize where experience and local communities' needs must be heard directly, not just translated, aggregated, or theorized by experts. Further, the array of relationships between universities and local communities demonstrate how outreach and reciprocal dialogue can be more effective at identifying and understanding social impacts and expectations regarding AI, beyond the privileged points of view and experiences of faculty, staff, administrators, and students. Universities need to amplify relevant community voices by providing a platform for responsive governance and participatory decision-making, and by building on their role as a knowledge commons (Hess and Ostrom 2003; Madison 2020). Technocratic governance—in which policies and standards reflect only the perspectives of experts, not of impacted individuals—presents a serious legitimacy problem in situations where AI is applied with direct impacts on the public, as with public services and in the financial industry. As academic researchers collaborate with local communities or engage in outreach, they might address this challenge via intentional engagement with feminist (e.g., Gurumurthy and Chami 2022; Hudson, Rönnblom & Teghtsoonian 2017) or Indigenous (e.g., Carroll, Rodriguez-Lonebear & Martinez 2019; Carroll et al. 2021; Tsosie 2019) models for governance, in which they utilize participatory approaches and draw on design justice principles (Costanza-Chock 2020). These perspectives will help ensure that an inequitable status quo is not simply reproduced around AI technologies, but instead that those who have historically been overlooked or harmed have a say in what is appropriate. Universities ought to transcend their ivory towers to open dialogue on pressing sociotechnical issues, "finding people where they are" (Venkatasubramanian 2022) and using the prestige of their academic, intellectual, and institutional platforms to make those voices heard.

Audit

Academic researchers can play a key role in auditing processes, often on behalf of regulatory agencies, as independent experts to scrutinize security, privacy, discrimination, and other sociopolitical concerns in complex systems. Given the fast pace of technological change, AI regulation should not be a fixed form of control; rather, continuous monitoring and feedback should update regulatory regimes. This monitoring process would resemble post-market

surveillance for drug safety and efficacy by the Food and Drug Administration. There are also incentives for industry to participate willingly, given the opportunity to mitigate costly harms and identify new opportunities via outside perspectives. This work, however, requires a robust and independent auditing system for AI technologies. Universities can support these activities by reducing red tape around collaboration with regulatory agencies and by providing training and resources to support appropriate reporting.

Advise on best practices

The three traditional legs of academic work—research, teaching, and service—all have a role in advising on best practices for AI. Indeed, outside critiques from academia are often very useful tools for technology industry insiders to draw upon in effecting change. Outside expertise is used extensively by government regulatory bodies, such as design justice critiques of (1) the use of facial recognition technologies, automated license plate readers, and other forms of data-driven policing and electronic monitoring by law enforcement in U.S. municipalities (Detroit Community Technology Project 2019; Hussain and Schwartz 2021; Kilgore 2015; Stop LAPD Spying 2021), and (2) airport screening protocols and technologies (Costanza-Chock 2020), for which outside critiques have informed new gender-neutral standards and changes to the Advanced Imaging Technology (AIT) being used (TSA 2022). Technological models and theories developed by scholars also often influence civil society via inspirational objectives; for example, fairness, accountability, transparency, and ethics (FATE) scholarship impacted critiques of Europe's AI Act by advocacy groups such as EPIC and AlgorithmWatch (Lomas 2021), in that those groups perceived that act as falling short of meaningful fairness and accountability measures. In developing advice on best practices, university researchers need not develop novel ideas, but rather should organize and integrate existing knowledge via systematization (Kitamura and Mizoguchi 2004). To effectively communicate systematized best practices, new methods of engagement may be needed to capture the attention of relevant parties, shifting the focus from curriculum and pedagogy to engagement. In public service, a renewed outward focus is warranted. Systematization as a form of research, engagement as a key to teaching, and outward focus in public service are kinds of work that should be rewarded in university incentive structures.

Conclusions

From advances in healthcare to applications that impact modern society and governance, the ubiquity and impact of artificial intelligence (AI) are growing. The rapid rate of AI development and deployment brings a significant risk of unintended societal consequences, most of which remain either insufficiently analyzed or obscured.

Among the emerging themes are new opportunities for academic and industry collaborations to define a governance framework for AI that ensures that stakeholders have a voice and are empowered to mitigate the adverse impacts that can emerge from AI technologies. Toward that end, a common set of principles and values should be adopted by academia and industry, to create common ground around research practice and to help develop new avenues to engage society. For instance, we should recognize that data are collected, and research is conducted, in a political and historical context, which can produce inequitable benefits and harms. The harms of AI technologies can disproportionately affect marginalized communities.

Universities can play a crucial role by building networks among policymakers, government, courts, people and communities, media, and industry. Importantly, universities should avoid preaching from the ivory tower and instead engage with stakeholders. For example, universities should extend the mandate for AI literacy beyond computing departments, engage with the broader community to ensure AI literacy, and thus enable comprehensive and critical engagement. To be effective, such coordination efforts and societal calls to action should be responsive to community and social needs, rather than be data- or profit-driven alone. Universities can further accelerate progress by rewarding work on the topic, such as by valuing the labor needed to build genuinely reciprocal partnerships with community members, and by rewarding efforts to incorporate social responsibility into academic courses. Such incentives can have a significant effect, particularly for early-career academics. Such work is also essential in educating students who can meaningfully engage in AI governance to ensure a more responsible AI future.

Acknowledgments

We are grateful to Mary Gray, David Kaiser, Kush Varshney, and Suresh Venkatasubramanian for speaking at the symposium. We thank Lisa Bievenue, Timothy Bretl, Julie Munoz-Najar, Klara Nahrstedt, Hari Sundaram, and Ann Witmer for reviewing previous drafts of this paper. Finally, we thank Jenny Applequist for editing the text carefully.

References

Baker, Ryan S., and Aaron Hawn. 2021. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-021-00285-9

Banaji, Shakuntala, Ram Bhat, Anushi Agarwal, Nihal Passanha, and Mukti Sadhana Pravin. 2019. WhatsApp vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India. London School of Economics and Political Science. Department of Media and Communications. https://www.lse.ac.uk/media-and-communications/assets/documents/research/projects/WhatsApp-Misinformation-Report.pdf

Battle-Baptiste, Whitney and Britt Russert, eds. 2018. *W. E. B. Du Bois's data portraits: visualizing Black America*. New York: Princeton Architectural Press.

Benjamin, Ruha, ed. 2019a. *Captivating technology: Race, carceral technoscience, and liberatory imagination in everyday life*. Durham, NC: Duke University Press.

Benjamin, Ruha. 2019b. *Race after technology: Abolitionist tools for the New Jim Code*. Medford, MA: Polity Press.

Brock, André, Jr. 2020. *Distributed Blackness: African American cybercultures*. New York: NYU Press.

Broussard, Meredith. 2018. *Artificial unintelligence: How computers misunderstand the world*. Cambridge, MA: MIT Press.

Brown, Nicole, Ruby Mendenhall, Michael Black, Mark Van Moer, Assata Zerai, and Karen Flynn. 2016. Mechanized margin to digitized center: Black feminism's contributions to combatting

erasure within the digital humanities. *International Journal of Humanities and Arts Computing* 10 (1): 110–25.

Brown, Nicole, Ruby Mendenhall, Michael Black, Mark Van Moer, Ismini Lourentzou, Karen Flynn, Malaika McKee, Assata Zerai, and Chengxiang Zhai. 2019. In search of Zora/When metadata isn't enough: Rescuing the experiences of Black women through statistical modeling. *Journal of Library Metadata* 19 (3–4): 141–62.

Brown, Steven E. 2008. Breaking barriers: The pioneering disability students services program at the University of Illinois: 1948–1960. In *The history of discrimination in U.S. education*, ed. Eileen H. Tamura, 165–92. New York: Palgrave McMillian.

Browne, Simone. 2015. *Dark matters: On the surveillance of Blackness*. Durham, NC: Duke University Press.

Buiten, Miriam C. 2019. Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation* 10 (1): 41–59.

Buolamwini, Joy, and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81:77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Burton, Emanuelle, Judy Goldsmith, Sven Koenig, Benjamin Kuipers, Nicholas Mattei, and Toby Walsh. 2017. Ethical considerations in artificial intelligence courses. *AI Magazine* 38 (2): 22–34. https://doi.org/10.1609/aimag.v38i2.2731

C/S2ESC - Software & Systems Engineering Standards Committee. 2021. *IEEE 7000-2021: IEEE standard model process for addressing ethical concerns during system design*. IEEE. https://standards.ieee.org/ieee/7000/6781/

Campolo, Alex, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. 2017. *AI Now 2017 report*. https://ainowinstitute.org/AI_Now_2017_Report.pdf

Carroll, Stephanie Russo, Desi Rodriguez-Lonebear, and Andrew Martinez. 2019. Indigenous data governance: Strategies from United States native nations. *Data Science Journal* 18. https://doi.org/10.5334%2Fdsj-2019-031

Carroll, Stephanie Russo, Edit Herczog, Maui Hudson, Keith Russell, and Shelley Stall. 2021. Operationalizing the CARE and FAIR principles for Indigenous data futures. *Scientific Data* 8. https://doi.org/10.1038/s41597-021-00892-0

Carroll, S. R., I. Garba, O. L. Figueroa-Rodríguez, J. Holbrook, R. Lovett, S. Materechera, M. Parsons, K. Raseroka, D. Rodriguez-Lonebear, R. Rowe, R. Sara, J. D. Walker, J. Anderson, and M. Hudson. 2020. The CARE principles for Indigenous data governance. *Data Science Journal*, 19 (1). http://doi.org/10.5334/dsj-2020-043

Cath, Corinne, Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, and Luciano Floridi. 2018. Artificial intelligence and the "good society": The US, EU, and UK approach. *Science and Engineering Ethics* 24 (2): 505–28.

Chan, Anita and Patricia Garcia. (Forthcoming.) Community data. In *The SAGE handbook of data and society: An interdisciplinary reader in critical data studies*, ed. Tommaso Venturini, Amelia Acker, Jean-Christophe Plantin, and Antonia Walford. London: Sage Publishers.

Chan, Anita. 2021. Community data clinic: Data methods from below. Paper presented at the American Studies Association Conference, 12 October, at San Juan, Puerto Rico.

Chan, Anita. 2020. Feminist data futures and the relational infrastructures of research. Paper presented in the Science and Technology Studies Seminar Series, 23 October, at Virginia Tech University.

Chan, Anita Say. 2014. *Networking peripheries: Technological futures and the myth of digital universalism*. Cambridge, MA: MIT Press.

Chun, Wendy. 2021. *Discriminating data: Correlation, neighborhoods, and the new politics of recognition*. Cambridge, MA: MIT Press.

Cifor, Marika, Patricia Garcia, T. L. Cowan, Jasmine Rault, Tonia Sutherland, Anita Chan, Jennifer Rode, Anna Lauren Hoffmann, Niloufar Salehi, and Lisa Nakamura. 2019. Feminist data manifest-no. https://www.manifestno.com/

Clearview AI, Inc. 2022. https://www.clearview.ai

Costanza-Chock, Sasha. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Couldry, Nick, and Ulises Mejias. 2019. T*he costs of connection: How data is colonizing human life and appropriating It for capitalism*. Palo Alto, CA: Stanford University Press.

Crawford, Kate. 2017. The trouble with bias. *Revolutions* [blog].18 December. https://blog.revolutionanalytics.com/2017/12/the-trouble-with-bias-by-kate-crawford.html

Crawford, Kate. 2021. *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. New Haven, CT: Yale University Press.

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data feminism*. MIT Press.

Dafoe, Allan. 2018. *AI governance: A research agenda*, v1.0. 27 August. Oxford, UK: Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf

Darnault, Cécilia, Titouan Parcollet, and Mohamed Morchid. 2019. Artificial intelligence: a tale of social responsibility. Association for the Advancement of Artificial Intelligence. https://hal.archives-ouvertes.fr/hal-02270410/document

Davis, Jenny. 2022. Consultation, collaboration, and consent: Research ethics and Indigenous methodologies for working in language archives and databases. Keynote address presented at the annual meeting of the Society for the Study of Indigenous Languages of the Americas (SSILA). 12 January.

Davis, Jenny L. 2017. Resisting rhetorics of language endangerment: Reclamation through Indigenous language survivance. *Language Documentation and Description* 14: 37–58. https://doi.org/10.25894/ldd147

Detroit Community Technology Project. 2019. A Critical Summary of Detroit's Project Green Light and its Greater Context. https://detroitcommunitytech.org/?q=content/critical-summary-detroit%E2%80%99s-project-green-light-and-its-greater-context

Dunne, Anthony, and Fiona Raby. 2013. *Speculative everything: Design, fiction, and social dreaming*. MIT Press.

Electronic Frontier Foundation. 2022. Community activists reach settlement with Marin County Sheriff for unlawfully sharing drivers' locations with out-of-state and federal agencies. Press release. 1 June. https://www.eff.org/press/releases/community-activists-reach-settlement-marin-county-sheriff-unlawfully-sharing-drivers

Eslami, Motahhare, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. Communicating algorithmic process in online behavioral advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery. https://doi.org/10.1145/3173574.3174006

Eubanks, Virginia. 2019. *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: Picador.

Eubanks, Virginia. 2011. *Digital dead end: Fighting for social justice in the information age*. Cambridge, MA: MIT Press.

Fiesler, Casey. 2019. Ethical considerations for research involving (speculative) public data. *Proceedings of the ACM on Human-Computer Interaction* 3 (GROUP): article no. 249. https://doi.org/10.1145/3370271

Fiesler, Casey, Natalie Garrett, and Nathan Beard. 2020. What do we teach when we teach tech ethics? A syllabi analysis. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 289–95. New York: Association for Computing Machinery. https://doi.org/10.1145/3328778.3366825

Flores, Michelle P., Lisa de la Rue, Helen A. Neville, Maria Valgoi, James Brooks, Een Sul Lee, Rebecca Ginsburg, Sammy Santiago, Kemuyah Ben Rakemayahu, Robert Garite, and Michael Brawn. 2014. Developing social justice competencies: A consultation training approach. *The Counseling Psychologist*, 42 (7): 998–1020.

Gentile, Mary C. 2010. *Giving voice to values: How to speak your mind when you know what's right*. New Haven: Yale University Press.

Ginsburg, Rebecca, ed. 2019. *Critical perspectives on teaching in prison: Students and instructors on pedagogy behind the wall.* New York: Routledge.

Goldsmith, Judy, Emanuelle Burton, David M. Dueber, Beth Goldstein, Shannon Sampson, and Michael D. Toland. 2020. Assessing ethical thinking about AI. In *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (9): 13525–8. https://doi.org/10.1609/aaai.v34i09.7075

Grosz, Barbara J., David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo. 2019. Embedded EthiCS: Integrating ethics across CS education. *Communications of the ACM* 62 (8): 54–61. https://doi.org/10.1145/3330794

Gurumurthy, Anita, and Nandini Chami. 2022. Beyond data bodies: New directions for a feminist theory of data sovereignty. Working paper 24. Data Governance Network. https://datagovernance.org/files/research/1641877014.pdf

Hedayati-Mehdiabadi, Amir. 2022. How do computer science students make decisions in ethical situations? Implications for teaching computing ethics based on a grounded theory study. *ACM Transactions on Computing Education* 22 (3): article 37. https://doi.org/10.1145/3483841

Hermann, Erik. 2021. Artificial intelligence and mass personalization of communication content: An ethical and literacy perspective. *New Media & Society* 24 (5). https://doi.org/10.1177/14614448211022702

Hess, Charlotte, and Elinor Ostrom. 2003. Ideas, artifacts, and facilities: Information as a common-pool resource. *Law and Contemporary Problems* 66 (1/2): 111–45.

Hudson, Christine, Malin Rönnblom, and Katherine Teghtsoonian. 2017. *Gender, governance and feminist analysis: Missing in action?* London: Routledge.

Hussain, Saira and Schwartz, Adam. 2021. EFF files new lawsuit against California sheriff for sharing ALPR data with ICE and CBP. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2021/10/eff-files-new-lawsuit-against-california-sheriff-sharing-alpr-data-ice-and-cbp

ImageNet [dataset]. 2021. https://www.image-net.org

Institute of Continuing Education, Cambridge University. 2022. MSt in AI Ethics and Society. https://www.ice.cam.ac.uk/course/mst-artificial-intelligence-ethics-and-society-2022

Irani, Lilly. 2021. Claiming democracy over digital infrastructures [video]. Just Infrastructures seminar series. 28 April. https://just-infras.illinois.edu/lilly-irani/

Jefferson, Brian. 2020. *Digitize and punish: Racial criminalization in the digital age*. Minneapolis, MN: University of Minnesota Press.

Jung, Minseok, and Madelyn Rose Sanfilippo. 2022. Mapping geographical biases of AI principles. Poster presentation at the iConference 2022. https://hdl.handle.net/2142/113756

Kandlhofer, Martin, Gerald Steinbauer, Sabine Hirschmugl-Gaisch, and Petra Huber. 2016. Artificial intelligence and computer science in education: From kindergarten to university. In *Proceedings of the 2016 IEEE Frontiers in Education Conference (FIE)*. IEEE. https://doi.org/10.1109/FIE.2016.7757570

Kang, Jian, Tiankai Xie, Xintao Wu, Ross Maciejewski, and Hanghang Tong. 2021. MultiFair: Multi-group fairness in machine learning. arXiv preprint arXiv:2105.11069.

Khisamova, Z. I., Ildar R. Begishev, and Ramil R. Gaifutdinov. 2019. On methods to legal regulation of artificial intelligence in the world. *International Journal of Innovative Technology and Exploring Engineering* 9 (1): 5159–62.

Kilgore, James. 2015. *Electronic monitoring is not the answer: Critical reflections on a flawed alternative*. Urbana-Champaign Independent Media Center. https://mediajustice.org/wp-content/uploads/2015/10/EM-Report-Kilgore-final-draft-10-4-15.pdf

Kirkpatrick, Keith. 2016. Battling algorithmic bias: How do we ensure algorithms treat us fairly? *Communications of the ACM* 59 (10): 16–7. https://doi.org/10.1145/2983270

Kitamura, Yoshinobu, and Riichiro Mizoguchi. 2004. Ontology-based systematization of functional knowledge. *Journal of Engineering Design* 15 (4): 327–51.

Krkač, Kristijan, and Ivana Bračević. 2020. Artificial intelligence and social responsibility. In *The Palgrave Handbook of Corporate Social Responsibility*, ed. David Crowther and Shahla Seifi. Palgrave Macmillan Cham. https://doi.org/10.1007/978-3-030-22438-7

Lessig, Lawrence. 1999. *Code: And other laws of cyberspace*. New York: Basic Books.

Lewis, Tamika, Seeta Peña Gangadharan, Mariella Saba, and Tawana Petty. 2018. *Digital defense playbook: Community power tools for reclaiming data*. Detroit: Our Data Bodies.

Littman, Michael L., Ifeoma Ajunwa, Guy Berger, Craig Boutilier, Morgan Currie, Finale Doshi-Velez, Gillian Hadfield, Michael C. Horowitz, Charles Isbell, Hiroaki Kitano, Karen Levy, Terah Lyons, Melanie Mitchell, Julie Shah, Steven Sloman, Shannon Vallor, and Toby Walsh. 2021. *Gathering strength, gathering storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel report*. Stanford, CA: Stanford University. http://ai100.stanford.edu/2021-report

Lomas, Natasha. 2021. Europe's AI Act falls far short on protecting fundamental rights, civil society groups warn. *Tech Crunch*, 30 November. https://techcrunch.com/2021/11/30/eu-ai-act-civil-society-recommendations/

Loui, Michael C. 2009. What can students learn in an extended role-play simulation on technology and society? *Bulletin of Science, Technology & Society* 29 (1): 37–47. https://doi.org/10.1177/0270467608328710

Ma, Alexandra, and Ben Gilbert. 2019. Facebook understood how dangerous the Trump-linked data firm Cambridge Analytica could be much earlier than it previously said. Here's everything that's happened up until now. *Business Insider*. 23 August. https://www.businessinsider.com/cambridge-analytica-a-guide-to-the-trump-linked-data-firm-that-harvested-50-million-facebook-profiles-2018-3

Madison, Michael J. 2020. Data governance and the emerging university. In *Research Handbook on Intellectual Property and Technology Transfer*, ed. Jacob H. Rooksby, 364–390. Cheltenham, UK: Edward Elgar Publishing.

McIlwain, Charton D. 2020. *Black software, the Internet & racial justice, from the AfroNet to Black Lives Matter*. New York: Oxford University Press.

Mendenhall, Ruby, Nicole Brown, and Michael Black. 2017a. The potential of big data in rescuing and recovering Black women's contributions to the Du Bois-Atlanta School and to American sociology. *Ethnic and Racial Studies* 40 (8): 1231–3.

Mendenhall, Ruby, Taylor-Imani A. Linear, Malaika McKee, Nicole Lamers, and Michel Mouawad. 2017b. Chicago African American mothers' power of resistance: Designing spaces of hope in global contexts. In *The Power of Resistance: Culture, Ideology and Social Reproduction in Global Contexts*, ed. R. Elmesky, C.C. Yeakey, and O. Marcucci, 409–28. Bingley, UK: Emerald Publishing.

Milner, Yeshimabeit. 2020. Abolish big data [video]. Databite no. 129. Data & Society Research Institute. 4 March. https://datasociety.net/library/abolish-big-data/

MIT Schwarzman College of Computing. 2022. Social and Ethical Responsibilities of Computing (SERC). https://computing.mit.edu/cross-cutting/social-and-ethical-responsibilities-of-computing/

Moon, Margaret R. 2009. The history and role of institutional review boards: A useful tension. *AMA Journal of Ethics* 11 (4): 311–6.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. *The Belmont report*. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html#xjust

National Institute of Standards and Technology (NIST). 2022. AI risk management framework: Initial draft. 17 March. https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf

Nelson, Alondra. 2013. *Body and soul: The Black Panther Party and the fight against medical discrimination*. Minneapolis, MN: University of Minnesota Press.

Noble, Safiya. 2018. *Algorithms of oppression: How search engines enforce racism*. New York: NYU Press.

O'Neill, Cathy. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown.

Online Ethics Center. 2022. https://onlineethics.org/

Partnership on AI (PAI). 2022. About us: Advancing positive outcomes for people and society. https://partnershiponai.org/about/

Petty, Tawana. 2018. *Towards humanity: Shifting the culture of anti-racism organizing*. Self-published.

Precarity Lab. 2020. *Technoprecarious*. Cambridge, MA: MIT Press.

Reagan, Leslie J. 2017. Timothy Nugent: "Wheelchair students" and the creation of the most accessible campus in the world. In *The University of Illinois: Engine of innovation*, ed. F. E. Hoxie, 50–9. Champaign, IL: University of Illinois Press.

Santoni de Sio, Filippo, and Giulio Mecacci. 2021. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology* 34; 1057–84. https://doi.org/10.1007/s13347-021-00450-x

Saveliev, Anton, and Denis Zhurenkov. 2020. Artificial intelligence and social responsibility: The case of the artificial intelligence strategies in the United States, Russia, and China. *Kybernetes* 50 (3). https://www.emerald.com/insight/content/doi/10.1108/K-01-2020-0060/full/html

Social Science One. 2022. https://socialscience.one/

Stop LAPD Spying Coalition. 2021. *Automating banishment: The surveillance and policing of looted land*. https://automatingbanishment.org/assets/AUTOMATING-BANISHMENT.pdf

Thaler, Richard H., Cass R. Sunstein, and John P. Balz. 2013. Choice architecture. *The behavioral foundations of public policy*, ed. E. Shafir, 428–39. Princeton, NJ: Princeton University Press.

Timberg, Craig. 2021, September 10. Facebook made big mistake in data it provided to researchers, undermining academic work. *The Washington Post*. https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists/

Transportation Security Administration (TSA). 2022. Transgender / non binary / gender nonconforming passengers. 31 March. https://www.tsa.gov/transgender-passengers

Tsosie, Rebecca. 2019. Tribal data governance and informational privacy: Constructing "Indigenous data sovereignty." *Montana Law Review* 80 (2): 229–268.

United Nations Human Rights Council. 2018. Report of the detailed findings of the Independent International Fact-Finding Mission on Myanmar. 39th session of the Human Rights Council, 28 September, agenda item 4. https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_CRP.2.pdf

Varshney, Kush. 2022. How can universities collaborate with external organizations and local communities to address questions of bias and discrimination in AI enabled technologies? Paper presented at the Symposium on Artificial Intelligence and Social Responsibility, 23 March, at the University of Illinois Urbana-Champaign.

Varshney, L. R., N. S. Keskar, and R. Socher. 2019. Pretrained AI models: performativity, mobility, and change. arXiv:1909.03290.

Venkatasubramanian, Suresh. 2022. AI governance: The role of the university. Paper presented at the Symposium on AI and Social Responsibility, 23 March, at the University of Illinois Urbana-Champaign.

Wiggers, Kyle. 2020. Researchers show that computer vision algorithms pretrained on ImageNet exhibit multiple, distressing biases. *VentureBeat*, 3 November. https://venturebeat.com/business/researchers-show-that-computer-vision-algorithms-pretrained-on-imagenet-exhibit-multiple-distressing-biases/

Williamson, Oliver E. 1996. *The mechanisms of governance*. New York: Oxford University Press.

Wu, Tim. 2018. *The curse of bigness: Antitrust in the new Gilded Age*. New York: Columbia Global Reports.

Yara, Olena, Anatoliy Brazheyev, Liudmyla Golovko, and Viktoriia Bashkatova. 2021. Legal regulation of the use of artificial intelligence: Problems and development prospects. *European Journal of Sustainable Development* 10 (1): 281-289.

Zimmer, Michael. 2020. 'But the data is already public': On the ethics of research in Facebook. In *The Ethics of Information Technologies*, ed. Keith Miller and Mariarosaria Taddeo, 229–41. Abingdon, UK: Routledge.

Zuboff, Shoshana. 2019. *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York: Public Affairs.