Responsible AI and Tech Justice: Curricular Examples on Racist Facial Recognition

Marie K. Heath
Loyola University Maryland

Daniel G. Krutka
University of North Texas

Shana V. White
Kapor Foundation

The release of ChatGPT in November of 2022 immediately resulted in concerns about the role of generative Artificial Intelligence in educational spaces. Many of the initial responses focused narrowly on either how to use AI as an educational tool or how to prevent students from cheating with AI. However, these reactive approaches fail to address many of the systemic social and educational issues that for-profit AI companies exacerbate, amplify, or create through their code.

In response, the Kapor Foundation (https://www.kaporcenter.org/) based out of the United States sought to ensure that the role of AI in educational spaces attended to issues of justice. Lead authors Shana V. White, Dr. Allison Scott, and Dr. Sonia Koshy worked with a team of authors and advisors to release the *Responsible AI and Tech Justice: A Guide for K-12 Education* framework in January of 2024. The purpose of this report is described as follows:

> The Responsible AI and Tech Justice Guide for K-12 Education is intended to articulate a new approach to teaching and learning in the era of rapid AI development and deployment, which prioritizes the interrogation of ethics, equity, and justice in the creation, deployment, and utilization of AI technologies and inspires the design and adoption of more equitable and just products and solutions. We anticipate that this guide and its utilization will continue to evolve and be refined over time. (p. 2)

They particularly emphasize the importance of this framework for computing education, but it is applicable to other areas of study as well. The authors utilize a racial and social justice lens to encourage students "to critically interrogate the ethical and equitable development, deployment, and impacts of AI, while simultaneously challenging, disrupting, and remedying the harms that these various technologies can cause within individual's lives, communities, and society at large" (p. 3).

The report includes six core components that can be taught explicitly or integrated in and across curriculum. In this section, we will share these core components and provide illustrative examples of how they may be taken up in various school spaces.

Responsible AI and Tech Justice in Practice

In an effort to offer illustrative examples of how to advance responsible AI and tech justice, we turn to Dr. Joy Buolamwini's 2023 book, *Unmasking AI: My Mission to Protect What is Human in a World of Machines.* We draw on this book because it is accessibly written for a popular audience and demands AI be responsible and just technology. Dr. Buolamwini rose to notoriety when trying to create an "Aspire Mirror" as part of a master's class at MIT. The mirror was designed to use facial recognition to track faces and show reflections of people who inspire the user. However, she realized that the programs she used were unable to accurately detect her darker skin. This led her to study facial recognition systems further. Her 2018 paper co-authored with Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," detailed the racial bias in facial recognition algorithms from companies with commercial AI products such as Amazon, IBM, and Microsoft. In response, some of these companies sought to address the racism coded into their technologies. These events garnered media attention, were detailed in the 2020 documentary *Coded Bias*, and also detailed in her book.

We will offer narrative examples from Dr. Buolamwini's work for each core component so educators might imagine what the curriculum could look like in their settings. We will also highlight other scholars and resources—many of which are cited in the Kapor Foundation's report appendix—to offer further examples of how the framework can be taught in schools. We italicize some key concepts or processes that educators might define and investigate with students.

Core Component 1: Examine the AI technology creation ecosystem from who designs and develops products and how they are developed, to who invests in their creation and benefits from their adoption.

The first core component invites educators and students to inquire into the full scope of AI—from creation to design to deployment. While designing her Aspire Mirror, Dr. Buolamwini (2023) recognized that it matters *who* designs a product. She conducted *algorithmic audits* to uncover the specific ways that the processes of facial recognition softwares and their algorithmic design—from data set inclusion and exclusion decisions, to the human coding of data, through the *black box* of machine learning, and up to the outcomes of the *coded gaze* on minoritized people—are encoded with bias and power.

Buolamwini sought to know what data was included in the training set, and who the coders were that created and approved the software, which failed to recognize darker skin. She found that the data sets were trained on and by overwhelmingly lighter skinned men, who she jokingly termed the “pale males.” The absence of more darker skinned and female faces meant the algorithm never “learned” what these faces looked like. The lack of diversity in the training data set resulted in inaccurate software. A lack of diversity in coders furthered the inaccuracy, because the designers were also unable to detect the inaccuracies of the software. After all, their facial recognition technology worked for them.

This component moves beyond the design and development and also invites investigation into who, specifically, invests in the creation and benefits from the deployment of AI technologies. We return to Dr. Buloamwini’s work as an example of scholarship which pursues this approach. Dr. Buolamwini notified each company that had developed ineffective facial recognition IBM and Microsoft worked to correct their flawed data and design. While some companies may be willing to invest the capital and time required to develop more inclusive data sets, other companies, like Amazon, refused to acknowledge the influence of her work and publicly attempted to undermine her methodologies.

The Kapor Foundation’s framework offers a series of questions to facilitate inquiry into this component. The questions probe assumptions about techno-optimism, explore connections between the demographics of coders and users, and consider potential harms of surveillance on marginalized people. The framework also offers a series of resources to support students and educators as they explore the ecosystem around the creation of AI technologies.

Core Component 2: Interrogate the complex relationship between technology and human beings, including human-computer interaction and topics of values, ethics, privacy, & safety.

Technologies are often referred to as tools, calling to mind an item that a human could pick up, employ for its intended purpose, and then put it away in the cupboard until the next time it’s needed. However, tools shape human behavior, both while they are in use and when they are put away on the shelf. The old saying, “to hammer, everything looks like a nail” is true both when the hammer is in our hands, and when the hammer is accessible to us in the toolbox influencing the ways we think about the next building project. This Core Component invites educators and students to think of technology not as a tool that can be picked up and put down at will, but rather as a force that alters the environments in which we live, learn, and interact.

As Dr. Buolamwini points out, the algorithms powering our newer digital technologies are almost always with us, rarely turned off and stowed away. They mediate our experiences with the world and nudge us to act and think in ways aligned with the machine. Nicholas Carr (2011) elaborated on the ways algorithmically driven search engines (e.g., Google) has shaped the way we understand, access, and consume information, minimizing our attention spans and altering our understandings of what counts as knowledge. Shoshanna Zuboff (2019) argued that the machines take advantage of the personal data we create when we use the technologies in order to sell us items that the machine learning algorithm predicts we will want. She termed this *surveillance capitalism.* Using deeply personal data including geolocation, demographics, purchasing history, musical preferences, medical diagnoses, and biometric data including wake, sleep, and heartbeat, companies can advertise and sell us goods targeted directly to what companies determine we need.

Core Component 2 includes questions to help interrogate these practices, including inquiries designed to investigate how algorithms nudge human behavior, interrogate the ethics of data mining and surveillance, and explore the notion of consent within the context of generative AI and algorithms. A sample of these questions include, How do algorithms influence human behavior? Who trains algorithms, how are they trained, and what are they trained to optimize for? What is HCI (human-computer interaction) & what are human-centered approaches to AI technology design and development? The suggested resources reference the Civics of Technology lesson designed by Dan Krutka and Zack Seitz which explores Dr. Buolamwini's work in the documentary *Coded Bias*, titled *Who is Responsible for Discriminatory Design*.

Core Component 3: Explore the impacts and implications of AI technologies on society, including positive benefits, negative consequences, and the perpetuation of exclusion, marginalization, and inequality.

Technologies, particularly emergent ones, are often discussed around their intended purposes, which are often framed by the technology companies who seek profits. As the Kapor Foundation report notes:

> Yet, despite the potential for AI to identify breakthroughs in disease prevention and treatment; improve efficiency in communication through autocomplete, virtual assistants, and chatbots; reduce business costs and improve efficiency and worker productivity; and enable greater personalization to improve educational outcomes for students with disabilities and multilingual learners, advancements in AI are not without concern. (p. 3)

Even these examples which—on their face—seem overwhelmingly positive, there can be negative consequences that emerge over time. For example, does autocomplete standardize knowledge, or even amplify harmful searches? Do virtual assistants or chatbots decrease human interaction and ingenuity? Are we even sure efficiency is the value that needs to be emphasized in a fast-paced, information-rich world? How is worker productivity gauged in white collar, knowledge work sectors? Does AI result in disproportionate surveillance on blue collar workers or students? Is white mainstream English further standardized and amplified by AI?

Dr. Buolamwini's work identifies how better facial recognition AI can result in both the invisibility of dark skin and the hypervisibility of Black bodies. She describes the false arrest of

Detroit man Robert Williams at his home and in front of his family because of a false match. The inaccuracy of the software resulted in real lives harms that disproportionately will affect minoritized groups. She also discusses how tenants in Brooklyn challenged a landlord's efforts to install facial recognition entry systems in their buildings. The primarily Black and Brown residents felt the system represented an unnecessary effort to increase surveillance without meaningful security benefits.

The Kapor Foundation framework poses questions such as: How do institutional and structural injustices get enhanced or replicated by technology and AI tools? What communities have disproportionately been impacted by surveillance? How do AI technologies contribute to climate change? How can AI technologies improve education and healthcare outcomes? They also provide popular press articles to learn more, and activities such as the *Is this a Weapon of Math Destruction?* lesson developed by Jacob Pleasants as part of the Civics of Technology project (https://www.civicsoftechnology.org/wmd-lesson).

Core Component 4: Interrogate personal usage of AI technologies to become critical consumers of products and address misuse, exploitation, and safety concerns.

Often, when the media refers to “AI” they are talking about generative AI; however, AI in the form of machine learning and natural language processing models has been part of many individuals’ daily lives for years. For instance, Netflix and Spotify algorithms, social media news feeds, and sidebar advertisements on web browsers all rely on machine learning, using data about our past behaviors to make predictions about our future desires. Similarly, even before generative AI, natural language processing models have been suggesting the next best word in our email and text messaging apps, nudging us toward standard grammar usage, and shifting the tone of our writing depending on the audience.

This Core Component encourages educators to uncover the cost that AI in our daily lives, including machine learning, natural language processing, and generative AI, extracts from us. For instance, the algorithmic power of each of these processes relies on accessing significant quantities of personal information, often obtained through surreptitious data scraping. The result of ever more personalized internet and social media experience can lead to radicalization, disinformation, and misinformation. The questions and resources provided in the framework facilitate discussion and interrogation of these ideas. They include questions like, What rights do you have to your own data? What rights do people have to protect themselves from surveillance and facial recognition? What protections are in place for individuals to refuse certain
technology and AI tools?

Core Component 5: Build a critical lens in the collection, usage, analysis, interpretation, and reporting of data.

In chapter 9 titled "Crawling through Data," Buolamwini details the deeply social, historical, and cultural knowledge that is needed to understand how data sets inform AI. She discusses the difficulty of classifying people by race and gender, which if not done well, can embed biases. She contends that AI requires "sociotechnical research, which emphasizes that you cannot study machines created to analyze humans without also considering the social conditions and power relationships involved" (p. 94). Unfortunately, many AI researchers create machine learning models that are largely guessing on these demographic categories. Moreover, many AI companies collect data sets without consent of the users whose images, ideas, and creations to train their systems. Such issues identify research ethics issues around who decides whose data is and should be used. Dr. Buolamwini's detailed discussion of methodological and ethical issues can be instructive to students about what responsible AI and tech justice might look like.

The Kapor Foundation framework challenges educators and students to engage in such difficult issues by posing questions such as: What data sources are used for AI technologies and can these sets have biases? What is the difference between training data, validation data, and testing data in datasets used for AI technologies? How is the data for this AI technology obtained? What are responsible and ethical approaches to data collection and usage? What is data sovereignty? They also provide a range of popular press articles to learn about similar issues, and resources for the classroom. For example, they recommend a "Data Privacy Drag and Drop Exercise" created by Autumm Caines that challenges students to understand how identity, data collection, and data outcomes are related and result in disproportionate harms for minoritized groups in particular (https://technoethics.digciz.org/index.php/drag-and-drop/).

Core Component 6: Minimize, mitigate, and eliminate harm and injustice caused by AI technologies through both the responsible and ethical creation process and individuals' and collective right to refusal.

The final core component encourages educators and students to take action. Throughout her book, Dr. Buolamwini provides examples of how to confront the harms of AI and take action. This ranges from her creation of works of art to challenge AI narratives and center people who are likely to be harmed or targeted to testifying before Congress on the topic. However, nothing illustrates her activism more than her creation of the Algorithmic Justice League (AJL; https://www.ajl.org/). The AJL site describes their purpose as:

> The Algorithmic Justice League is an organization that combines art and research to illuminate the social implications and harms of artificial intelligence. AJL's mission is to raise public awareness about the impacts of AI, equip advocates with resources to bolster

> campaigns, build the voice and choice of the most impacted communities, and galvanize researchers, policymakers, and industry practitioners to prevent AI harms. (n.p.)

The AJL provides students an example of how they can be agents of change in demanding responsible AI and tech justice. In addition to providing a range of popular press sources on the topic, the Kapor Foundation framework also includes a computer science unit where students seek to answer the question, how can we fight algorithmic bias? On different days of the unit, students will investigate sources to answer questions such as: What is an algorithm? What is algorithm bias? Who does algorithm bias affect? What can be done about algorithm bias? What will I do to fight algorithm bias? What are others doing to fight algorithm bias? The lessons even include source material where students learn about the AJL.

Conclusion

One of the challenges of advancing responsible AI and tech justice in schools is helping students, educators, and community members see what is at stake. The Kapor Foundation's *Responsible AI and Tech Justice: A Guide for K-12 Education* offers a framework that can support educators, but it must be accompanied by stories that compel people to care. In this article, we recommend Joy Buolamwini's new book as an accessible and vivid example to which educators and students might turn to both define the problem and pursue solutions. There is much work to be done, and the Kapor Foundation report concludes by seeking feedback and collaboration from people from across educational spaces and disciplines. If AI is to be part of a humane and just world, educators can work alongside students to program that world.

References

Buolamwini, J. (2023). *Unmasking AI: my mission to protect what is human in a world of machines*. Random House.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). Proceedings of Machine Learning Research.

Carr, N. (2011). *The shallows: What the Internet is doing to our brains*. W.W. Norton & Company.

Kapor Foundation. (2024). *Responsible AI & Tech Justice Framework.* Kapor Foundation. https://kaporfoundation.org/wp-content/uploads/2024/01/FINAL-FULL-GUIDE-kapor-foundation-responsible-ai.pdf

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Routledge.