Authentic Integration of Ethics and AI through Sociotechnical, Problem-Based Learning

Ari Krakowski,[1] **Eric Greenwald,**[1] **Timothy Hurt,** [1] **Brandie Nonnecke,** [2] **Matthew Cannady** [1]

[1] Lawrence Hall of Science, University of California, Berkeley
[2] Center for Information Technology Research in the Interest of Society (CITRIS), University of California, Berkeley
(akrakowski, eric.greenwald, thurt, nonnecke, mcannady)@berkeley.edu

Abstract

Growing awareness of both the demand for artificial intelligence (AI) expertise and the societal impacts of AI systems has led to calls to integrate learning of ethics alongside learning of technical skills in AI courses and pathways. In this paper, we discuss our experiences developing and piloting the TechHive AI curriculum for high school youth that integrates AI ethics and technical learning. The design of the curriculum was guided by the following pedagogical goals: (1) to respond to the capacity-building need for critical sociotechnical competencies in AI workforce pathways; and (2) to broaden participation in AI pathways through intentional instructional design to center equity in learning experiences. We provide an overview of the 30-hour learning sequence's instructional design, and our "4D Framework," which we use as a heuristic to help students conceptualize and inspect AI systems. We then provide a focused description of one of three chapters that make up the sequence. Finally, we present evidence of promise from an exploratory study of TechHive AI with a small sample of students, and discuss insights from implementation, including from our use of established resources for AI learning within the learning sequence as well as those created by our team.

Introduction: The Need for an Integrated Approach

Artificial intelligence (AI) permeates all aspects of society: AI is in our phones and our TVs, in our thermostats and our cars, and it is becoming integral and indispensable to the practice of modern science and a growing number of industrial sectors. AI systems influence the news we read, our medical care, and our likelihood of being hired, approved for a loan, or arrested. AI is also increasingly integral to systems in the public sector and to the practice of modern science. AI-enabled systems have advanced climate change research through improved Earth system modeling (Huntingford et al. 2019); accelerated biomedical research to improve disease prevention, treatment and monitoring (Yu, Beam, and Kohane 2018); and have the potential to promote more equitable distribution of social services (Nonnecke et al. 2020). Decisions about where and how AI is applied reflect the values of those making the decisions, and the status quo in many of our institutions. Given the rapid uptake of AI across disciplines and workforce pathways, AI expertise is needed across economic sectors (Costello 2019; World Economic Forum 2020). At the same time, we face an acute shortage of workers with AI expertise (McKendrick 2020; Zwetsloot, Heston, and Arnold 2019), suggesting a need for expanded academic and career pathways to counter the present workforce shortage. Accordingly, there is growing recognition of the importance of building AI literacy with K-12 youth, and a number of research efforts in the past ten years have investigated approaches and provided guidelines aimed at incorporating AI concepts into curricula for K-12 youth (Heinze, Haase, and Higgins 2010; Long and Magerko 2020; Touretzky et al. 2019). While technical knowledge is necessary to build AI systems, it is insufficient for understanding who will be impacted by the system and how, or for considering why and whether an AI system should be built in the first place.

The broad need for more workers intersects with slow progress toward equitable participation in CS pathways (Scott et al. 2018), which in turn translates to a starkly homogeneous AI workforce (West, Whittaker, and Crawford 2019). Moreover, positioning a greater diversity of youth as technological innovators is particularly urgent given growing awareness of how bias can be embedded in technologies and their applications (Hajian, Bonchi, and Castillo 2016), replicating and perpetuating bias against women, people of color, and low-income people (Ali et al. 2019; Buolamwini and Gebru 2018; Eubanks 2018; Lambrecht and Tucker 2019). The narrow, disproportionate representation of voices in the design of AI systems contributes to concomitantly disproportionate harm, reinforcing — and conferring false legitimacy to — oppression and discrimination. To bring about responsible AI, there is a growing call from scholars to advance equitable participation in AI fields, and to center and amplify the perspectives of communities most impacted by AI harm (Hampton 2021; Tech Can't Fix This 2020; McLennan et al. 2020; West, Whittaker, and Crawford 2019). Thus, as AI systems promise to transform the future of work and present exciting possibilities for innovative deployments across society, the ubiquity of AI presents troubling challenges: "algorithmic bias" in AI models (Buolamwini and Gebru 2018; Hajian, Bonchi, and Castillo 2016) and ill-conceived deployments of AI risk cre-

ating and perpetuating socially detrimental outcomes, such as discriminatory bias in the judicial system (Angwin et al. 2016), recruitment and hiring practices (Dastin 2018), and health care delivery (Obermeyer et al. 2019).

However, as calls to attend to the ethical dimensions of AI increase (Garrett, Beard, and Fiesler 2020; Grosz et al. 2019), and ethics training is beginning to be integrated into data science and computer science curricula, technical and the ethical learning opportunities are still largely bifurcated and the most effective models for such integration are still unclear. While many computer science and artificial intelligence programs at universities offer, and sometimes require, ethics courses as part of their degree program, rarely are the technical courses taught in such a way as to integrate the ethical decisions with the technical decisions (Saltz et al. 2019). Similarly, there are an ever-growing number of resources for teaching AI at the K-12 level that have lessons and activities on AI ethics (e.g., MIT AI Ethics Education Curriculum, Code.org's Machine Learning and Bias Lesson). However, many of these resources separate the AI ethics learning from the technical AI learning in the form of distinct lessons or modules, and/or abstract the AI ethics from real-world contexts (i.e., classifying fish vs trash to learn bias doesn't necessarily help students understand discrimination). This presents two notable issues for AI learning: (1) the separation of AI technical and ethical learning communicates that ethical concerns and technical skills *can* be disentangled in the design of responsible AI systems; and (2) the decontextualization of ethical concerns from real-world consequences of AI may fail to support learners in developing understanding of responsible AI design. Additionally, perhaps as a result of the separation and decontextualization of technical and ethical AI learning, the ethical dimensions of AI models are relatively de-emphasized relative to AI technical skills across learning experiences with respect to instructional time, which can communicate that ethical issues are less important than technical skills.We argue that the critical need to direct the power of AI toward socially desirable outcomes and simultaneously forestall deleterious outcomes demands that AI curricula adopt a transdisciplinary approach that positions ethical issues as integral to and considered throughout AI development and deployment.

In this paper, we describe our preliminary work designing and implementing the TechHive AI curriculum, which engages high school youth in integrated, sociotechnical learning of AI technical and ethics literacies. Our work explores the integration of ethics training within a technically-focused curriculum for high school students. Our pedagogical goals guiding the design of the TechHive AI curriculum are twofold: (1) to address the need for transdisciplinary approaches to AI development where technologists and social scientists receive training across their fields to facilitate collaborative work; and (2) to encourage a diversity of students to pursue AI pathways by positioning sociotechnical analyses and perspectives as critical for responsible AI development and use. The TechHive AI high school curriculum joins recent efforts at the elementary (Kim et al. 2021) and middle school (Williams, Kaputsos, and Breazeal 2021) levels exploring how to integrate learning of ethical implications of AI with AI technical skills such that youth develop this critical sociotechnical knowledge.

The TechHive AI Curriculum

Pedagogical Approach

Drawing on the expertise of a transdisciplinary team, with expertise in K-12 learning design, AI, and AI ethics and policy, we developed and piloted the TechHive AI curriculum, a 30-hour informal learning program designed to engage high school youth in sociotechnical, transdisciplinary learning at the intersection of AI technical development and AI ethics. Transdisciplinary learning design integrates the learning of social systems approaches with AI technical approaches (Crawford and Calo 2016), positioning AI systems as objects of critical interrogation and evaluation against values from learners' communities and lived experiences. Such approaches elevate the importance of ethics in AI by positioning ethics principles as core to best practices for AI development and deployment. To support this in practice, we built upon the existing *Framework for Responsible AI* (Nonnecke 2018) and introduced an expanded version to students as a heuristic through which they could interrogate AI systems. This *4-D Framework for Responsible AI* encourages youth to evaluate the ethical considerations of an AI system by asking them to:

1. Make a **Determination** about whether AI is an appropriate tool for the defined task;
2. Question the **Data** being used in the AI System;
3. Consider the affordances and limitations of the AI system's **Design**; and
4. Consider how the AI system's **Decision** will impact real world systems.

We developed the materials using a problem-based learning (PBL) approach. Rooted in the educational learning theories of experiential learning (Kolb 2014), constructivism (Bransford et al. 2000), and situated learning (Greeno 1998; Lave and Wenger 1991), PBL employs real-world problems — often transdisciplinary, complex, and "messy" — as meaningful contexts to motivate learner-centered, collaborative knowledge construction and integration (Hmelo-Silver 2004; Savery 2015; Torp and Sage 1998). Through iterative design and implementation across two student cohorts, we identified and piloted three real-world AI PBL contexts that could enable developmentally appropriate learning of technical concepts: (1) college admissions, (2) health care, and (3) social media. These three PBL contexts also lent themselves to a progressive build for learners to develop understanding of and facility with AI concepts and skills, including application of the 4-D Framework to operationalize the responsible AI principles that readily arise from each PBL context.

Reflecting the transdisciplinary model aimed at integrating sociotechnical AI conceptual learning, the overarching question motivating student engagement across the curriculum is: *How can we design effective and responsible AI systems?* The curriculum is structured into three sequential chapters organized around each central PBL context. Each

chapter begins by introducing youth to the PBL context that will motivate learning. Leveraging pedagogical strategies that elicit and incorporate students' funds of knowledge (Barton and Tan 2009; Moll et al. 1992; Verdin, Godwin, and Capobianco 2016), students learn about the rationale for the use of AI in each context, draw on their collective lived experience to interrogate the problem space that AI is being considered for, discuss the potential ethical concerns for each context, and get introduced to salient stakeholder positions to develop an initial understanding of how an AI system might be employed in the given PBL context. Building from this preliminary understanding, students engage in multimodal (Jewitt et al. 2001) instructional activities (e.g., unplugged activities, structured discourse, and media, alongside engagement with AI models), to provide rich opportunities and multiple access points for students to build increasingly complex understanding of concepts.

The instructional sequence was designed to support students to progressively integrate ideas into a coherent whole. We define coherence in relation to logical consistency and depth: student learning experiences should build logically, be motivated by questions about phenomena that activate students' prior knowledge, and enable integration of new knowledge to construct a deep understanding of the phenomenon (Fortus and Krajcik 2012). We worked to approach coherence from a student perspective (Reiser, Novak, and McGill 2017), designing learning experiences to connect to, and build upon, a student's emerging understanding, rather than relying on abstract connections to ideas or questions accessible only to those with a more expert understanding. To support coherence in instructional design, we employed a tool called a Coherence Flowchart (Amplify Science 2021) — a visual schematic of the curricular storyline (Ramsey 1993; Roth et al. 2011) that represents the flow of questions that motivate inquiry; activities that engage students in investigation, knowledge construction, and sensemaking; and (3) the focal concepts students are working toward understanding. In our design process, the Coherence Flowchart served as a critical tool by: 1) supporting explicit attention to coherence in the design of learning experiences and 2) serving as an object of shared inspection and accountability, against which proposed sequences of instruction could be systematically evaluated and iteratively revised.

Description of Chapter 1 Learning Experience

In the following section, we describe the flow of activities in the first chapter, focused on the context of AI models used in college admissions (see Figure 1, Chapter 1 Coherence Flowchart, that provides a high-level view of the storyline of the chapter). As with other TechHive AI PBL contexts, the college admissions context sought to ground instruction in the principle of "thick authenticity" (Shaffer and Resnick 1999) by positioning learning as personally relevant, meaningful beyond the classroom walls, and aligned with disciplinary practice. As high school youth, many of the students were already considering future college applications and the factors that determine whether or not their applications are deemed competitive. The learning experiences in the chapter were designed to build student understanding of how machine learning models are trained and evaluated. Students were given opportunities to inspect the *data* being used to train the models and change the *design* of the models through practices such as feature selection. Alongside these technical concepts, the learning sequence introduced students to the responsible AI principle of *fairness* as they investigated how the *data* and its incorporation into the *design* of AI systems can lead to biased *decisions* that mirror and perpetuate societal disparities. Concomitantly, students were also introduced to the responsible AI principle of *accountability* through activities focused on evaluating whether *decisions* of AI models align with desired outcomes and figuring out how to address ethical consequences that may arise from their deployment. As the first chapter in the instructional sequence, the AI models that students investigated were intentionally more transparent relative to more opaque AI models investigated in subsequent chapters to enable students to more readily interrogate and inspect them.

The first activity introduces students to the PBL context of AI in college admissions, and the first investigation question: *How is AI used in this context?* In this activity, students model the process of college applicant selection mediated by (human) admissions officers. Provided with 15 hypothetical applications, student groups were tasked with analyzing the information provided for each applicant to compare attributes across the applicant pool and ultimately select the five "best" applicants. Student groups shared their applicant selections, and discussed the process they used for selection. In both student cohorts, no two groups arrived at the same selection of applicants. The activity concluded with a reflective discussion, in which students noted the inconsistency in applicant selection decisions, as well as the time-consuming nature of the decision-making process. This activity was designed to support student understanding that AI systems can be developed to make decisions more quickly and more consistently than humans can (key concept 1, see Fig. 1). This activity was followed with a structured discourse activity (Activity 2) in which students closely examine and discuss the different types of applicant data available to human admissions officers and AI admission models to determine applicant selection decisions (building toward key concept 2, see Fig. 1). Students were introduced to the concept of feature selection wherein an AI developer has to choose which features of the available data they want the system to utilize when making a decision. Students discussed many features, like grade point average, enrollment in AP classes, extracurriculars, work experience, school rank, and student rank. The conversations were facilitated to touch on the point that, depending on which features are or aren't included, some preference may be given to applicants of a particular gender or ethnicity, or from certain socioeconomic backgrounds. As an example of this, students discussed how schools in wealthier neighborhoods offer more AP classes than schools in relatively less wealthy neighborhoods, noting that the selection of AP class completion as a feature to inform admission would result in decisions that skew towards students from wealthier schools. Students also shared their ideas about whether outcomes such as this are acceptable, and whether different features might yield different decisions.

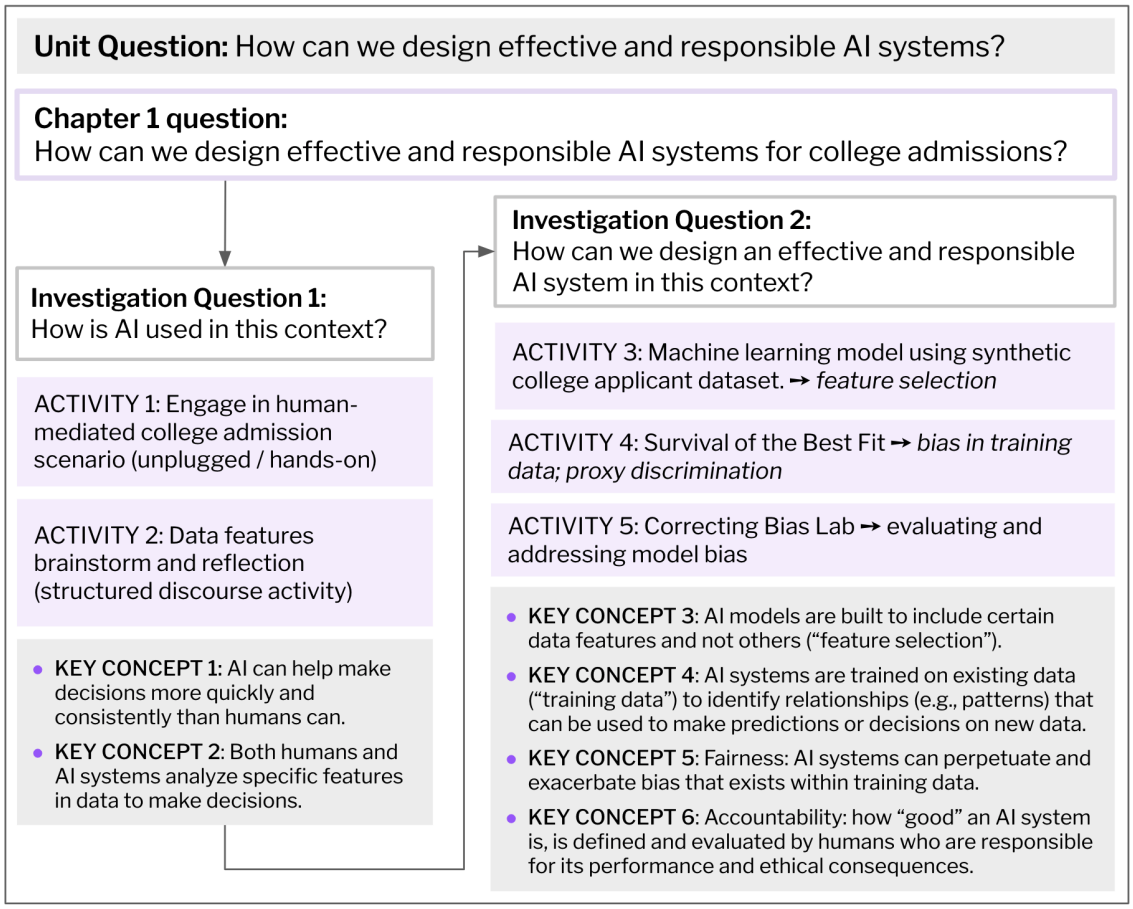

Figure 1: Coherence Flowchart for Chapter 1

Together, the first two activities motivate student inquiry about the second investigation question: *How can we design an effective and responsible AI system in this context?* Students began investigating this question by engaging with a machine learning model we developed in a Google Colaboratory notebook. To create the model, learning designers created a synthetic college applicant dataset and built a simple user interface through which students could explore different feature selection permutations, and evaluate the resulting model decisions. Students were provided with initial (fictionalized) historical applicant decision data that the college admissions office proposed as representative of decisions aligned with their goals for student recruitment. Students were readily able to identify feature selection permutations that maximized "accuracy" with respect to replicating historical applicant decision patterns. Student groups then interrogated the outcomes of their model and discussed whether they thought it was a "good" model beyond fidelity to patterns in historical admission data. As they explored additional feature selection permutations and evaluated their models, students discovered that many of the features that contributed to higher "accuracy" were also well-correlated with familial wealth or male gender, and thus produced biased admissions outcomes. Building on this activity, students next investigated an online interactive digital tool called Survival of the Best Fit (SOTBF) in which users inhabit the role of a hiring manager (which we position as analogous to that of college admissions officer) who is asked to use a machine learning model to automate the hiring process in order to evaluate an otherwise prohibitively high number of applicants. SOTBF walks users through a high-level machine learning model development process wherein applicants' data and data from a large company are used to create a hiring decision algorithm that, ultimately, is revealed to produce unwittingly biased hiring decisions. These activities culminated in critical discussions problematizing the notion of "accuracy" in machine learning and the risk that models trained to replicate historic data patterns can likewise replicate and perpetuate structural disparities "baked into" training data. Students also developed an understanding of how AI models can result in proxy discrimination (Prince and Schwarcz 2019), wherein a model can lead to discriminatory outcomes because it includes features that co-vary with protected characteristics, such as race or sex, even when those covariate features are excluded from the model.

Activities 3 and 4 led students to ask further questions about the responsible AI principle of *fairness*, and how

AI models can be designed to predict, understand, minimize, and continuously monitor bias in their outcomes. Motivated by these questions, the fifth activity engaged students in training a machine learning model to identify and mitigate bias by optimizing for fairness. Staying within the analogous context of hiring, students used a Colaboratory notebook from UC Berkeley's Daylight Lab (https://cltc.berkeley.edu/mlfailures), which contains synthetic data that was generated to replicate the well-documented gender pay gap that exists in many industries. Students trained a regression model to fit the data and then evaluated the model with respect to both how well it fit the data and whether it produced biased outcomes. After identifying pronounced gender bias, students were introduced to a new metric: a fairness metric (Chouldechova and Roth 2018; Gillen et al. 2018) wherein students determined a priori what a "fair" outcome would be (such as gender parity in hiring decisions), and then trained their model to optimize for that metric. Students then evaluated different models to see how well they performed with regards to both fairness and "accuracy" (i.e., applicants "hired" by the AI are likely to have been hired in real life). Optimizing for these two metrics simultaneously created a dilemma: as the model increasingly accounted for years of employment, the gender pay gap historically encoded in past hiring decisions became increasingly prominent; thus, the students had to figure out what degree of fairness they thought was appropriate for this model. This introduced the responsible AI principle of *accountability*, wherein those responsible for developing and implementing an AI-enabled tool should monitor and be accountable for its performance and ethical ramifications. In addition, confronting challenges of surfacing and mitigating biased outcomes produced by AI models in applicant selection decisions led students to discover that no technical "fix" can "solve" bias and other structural inequities, and that responsible AI design and deployment necessitates thoughtful *determination* as to whether AI should be used in a given context. Together, the last three activities in Chapter 1 were designed to support students in building understanding of key concepts 3 through 6 (see Fig. 1).

Implementation Insights

Context for Study

As summarized in Table 1, the TechHive AI instructional sequence was implemented with two cohorts of high-school-aged students, with some revisions to the curriculum between cohorts. Participants were recruited through our institutional networks, with a focus on recruiting from schools and communities with populations traditionally minoritized in STEM pathways. For both cohorts, learning experiences took place within the context of a global pandemic and were necessarily remote. While originally developed for in-person learning, with extensive opportunities for peer-peer collaboration and discussion, the instructional sequence was revised to provide collaborative experiences using web-based platforms, including Google Colaboratory notebooks, chat, and Zoom breakout rooms. The research team administered baseline and post-instruction surveys, and conducted individual interviews and focus groups with youth from the program, as summarized in Table 1. The conversations aimed to: (1) gauge students' confidence in applying the responsible AI principles and 4-D Framework to various scenarios; (2) understand their views in regards to technology and ethical concerns; and (3) determine if the students consider evaluating ethics in AI a pressing, possible, and productive need. We conducted focus groups with each cohort (4 participants in the first cohort and 10 participants in the second), but largely due to the complexities of administering surveys and conducting interviews in a remote context, we were only able to collect complete data sets (survey, focus group, and interview) from a subgroup of 4 students per cohort. For each cohort, the sample reflected the demographic diversity of the full cohort. To minimize selection-bias, each sample was evaluated by course instructors to ensure it reflected a mix of student engagement levels during instruction.

Analysis of data from pre/post surveys, individual interviews, and focus group interviews, offers insight into the extent that participating youth were able to successfully integrate technical concepts about AI systems with the principles of responsible AI. The data presented below are drawn from student responses to a series of scenario-based interview and survey prompts provided at the beginning of and then again immediately after the instructional sequence. The prompts asked students to evaluate a proposed AI system and identify areas of concern and/or questions to ask of the system design or its deployment. For example, one prompt asked about the use of facial recognition systems in criminal justice:

> *Your town/city is considering using facial recognition technologies for their police department. What would you want to know about it in order to decide if this is something they should use or not?*

This scenario-based approach to the interviews provided rich insight into how ideas from the learning sequence were being incorporated into youth's conceptualization of AI systems, and how those ideas get applied by youth to evaluate AI systems from multiple perspectives. For the second cohort, we refined and expanded the use of the scenarios to include them as open-ended prompts in the pre/post survey.

Initial Findings

As illustrated in the excerpts that follow, our analysis of individual and focus group interviews revealed increased sophistication and specificity in responses to similar prompts, between the beginning of instruction and the end of the program. The increased sophistication we observed came in the form of clear connections between the ethical concerns students raised and the technical decisions that impact those areas of concern (e.g., participants identified the need to use diverse datasets to train AI systems for criminal justice and college admission decisions, and recognized the particular importance of attending to false positives and false negatives when evaluating AI systems for healthcare). In the excerpted survey responses in Table 2, the student responds to the same prompt about the use of facial recognition systems for polic-

Cohort	Program Participants	Gender (self-report)	Ethnicity/Race (self-report)	Research Participants
Spring (3/2/21-5/8/21)	8 11th and 12th grade students	4 Male; 4 Female;	2 Asian; 1 Black/African-American; 1 Hispanic/Latino; 2 White; 2 identified as both Asian and White	4 focus group interview participants, all of whom completed surveys and individual interviews
Summer (7/19/21-7/30/21)	16 10th, 11th, and 12th grade students	11 Male; 5 Female;	7 Asian; 1 Black/African-American; 1 Hispanic/Latino; 1 Middle Eastern; 1 Native Hawaiian/Pacific Islander; 3 White; 1 identified as both Asian and White; 1 identified as both Hispanic and White	10 focus group interview participants, 4 of whom completed surveys and individual interviews

Table 1: Cohort Information

Pre Survey Response	Post Survey Response
I want to know if it registers all faces the same and if it has any bias. Hopefully, it is extensively researched, good, and unharmful to the people in my town.	*I would like to know what extent the technology is being used. Is it just for extra help or is that the main key for solving crimes? Is the technology going to have a fair and unbiased historical training or is it going to impact different people? Is the algorithm going to be a black box or is it transparent?*

Table 2: Pre & Post Survey Responses

ing prior to (left side) and just after (right side) participating in the learning sequence. While the student's initial response demonstrates an awareness of the problem of bias in facial recognition systems, that student provides considerably more detail in the post-instruction response, attending to questions about the role the AI system plays in the overall criminal justice space, the nature of the training data set, and whether the system is open to inspection.

We also see evidence in this excerpt that the complexity introduced by explicit attention to responsible AI was not impeding learning for the high school-aged youth, but may have in fact been helping to motivate it. For example, the student quoted in Table 2 was able to navigate questions of power dynamics (who decides), perspective and positionality (fair for whom and in what context), as well as appreciating how historical biases and inequities can become instantiated in AI systems (e.g., recognizing that admission/hiring screening systems work by identifying patterns in extant data, which may be the product of past inequities in who gets hired or accepted into a competitive college).

A primary goal of TechHive AI was to foster a sense of agency among students for their role in evaluating the design and deployment of AI systems. Indeed, by the end of the program, the students we interviewed all adopted a stance of investment in their own role of monitoring AI systems around these issues. In fact, each student we spoke with self-identified as a capable agent in this work, endorsing a perspective that they, and youth like them, have a critical role in evaluating AI systems and that the task is not something better left to "the experts."

> *I think it's really important [to be able to evaluate AI systems], especially for people my age since our generation is one growing up with a lot of technology. Because AI does seem to be a big part of the future of technology, I think it's important for people my age to understand how it works so that when we go into jobs in the future we can use AI systems and implement those into our jobs because I'm guessing they'll be implemented. So it's important that we know how they work and how to ensure that they are working properly.*

When we asked students what about the program contributed to their learning, they referenced the nesting of learning within relevant contexts such as college admissions and social media. Students also cited the rich group discussions that helped them integrate ethics and cybersecurity concerns with the technical features of an AI system. Further analyses to better understand how the design of instruction contributes to student learning are ongoing. We also saw evidence for nearly all students who participated in the research study that they found the topic, examining AI systems through the lens of ethical considerations, to be compelling and critical for their and their peers' future.

Discussion

A need for transdisciplinary understanding of AI systems. Given the ever-growing role of AI across disciplines, it is critical that all learners — whether or not they aspire to pursue academic or workforce pathways in AI — develop a foundational understanding not only of how AI systems operate, but also of the principles that can guide the responsible development, implementation, and monitoring of those systems. Incorporating responsible AI principles into AI development for the future workforce will demand that stu-

dents have the capacity to adopt different perspectives when evaluating AI systems and take into consideration potential ethical implications in the pursuit of desired technical outcomes. An approach that integrates AI technical and ethical domains should not be limited to those already advanced along academic and career pathways to AI. Moreover, as AI deployments expand from product recommendations to critical services like health care and education, AI systems are fast becoming integral to the functioning of society. Particularly in relation to these critical services, ethical concerns are paramount, and understanding how they manifest in AI systems helps consumers ask salient questions, evaluate options, and make informed decisions. Finally, it is important to recognize that we are at a policy crossroads with AI, and that we as a society are grappling with how best to regulate and prioritize public investment in and oversight of AI systems. In democracies, AI policies and governance are, and will be, "on the ballots," so it is important that there will be a diverse and educated electorate to actively engage with these policy concerns.

Within this broader context of AI's increasing role in society, we see promise in a transdisciplinary approach that is grounded in the complexities of specific problems for which AI systems might be developed and deployed. In this regard, the curriculum designers and educators observed that the problem-based learning contexts appeared to motivate student engagement in the informal learning context (i.e., participation in digital discourse through chat discussions; contributions to co-created ideation and reflection documents). The instructional sequence described here, and the underlying instructional design model aimed at coherent integration across sociotechnical domains, can support practitioners, learning designers, and researchers in engaging students in technical and responsible AI learning. The instructional design and 4-D Framework can complement international K-12 AI education efforts to articulate what students should understand about AI across grade bands (Touretzky et al. 2019), and how instructional materials and educators can support students in building that understanding (Greenwald, Leitner, and Wang 2021; Lee et al. 2021; Williams, Kaputsos, and Breazeal 2021).

A need for tools to support sociotechnical inquiry through experiential learning. Through the iterative design, development, and implementation pilots of TechHive AI, we saw considerable promise in learning activities in which technological resources enabled youth to engage deeply with both the technical and the ethical aspects of AI. In particular, an insight emerging from this work was the value of experiential learning (Kolb 2014) for building synergistic understanding of the technical concepts and responsible AI principles. That is, there is 'educative currency' in building knowledge about the ethical consequences of AI through designing, directly manipulating the inputs of, and interrogating the outputs of AI models. Notably, in the two activities described in the *Description of Curriculum* section that involved Colaboratory notebooks, students were able to work directly with data and machine learning models such that they could see the direct impacts their design decisions made on evaluation metrics like fairness. Conversely, when digital resources that enabled cohesive interplay between technical and ethics learning were unavailable or not well-matched to learning goals, we struggled to encourage students to cognitively integrate the technical and ethical considerations. For example, in the chapter focused on the use of natural language processing for content generation on social media platforms, students investigated a word-embedding visualization tool, Embedding Projector, to build an understanding about how word embeddings, learned from corpuses of natural language, can embed biases (Bolukbasi et al. 2016). Yet, students' interaction with this tool was abstracted from the PBL context, which appeared to hinder students' ability to make informed predictions about how such embedded biases could lead to ethical consequences on social media platforms, in contrast to their ability to make such predictions about unfairness in hiring (discussed above). These observations align with broadly held experiential and constructivist theories about how people learn—through cycles of investigation, collaborative sensemaking, and application—to encourage abstraction of underlying concepts.

As AI becomes embedded within our political, social, and economic institutions, mitigating bias, discrimination, and threats to public safety are paramount. It is therefore critical that AI systems are built with attention to responsible AI principles and practices. A sociotechnical curriculum may equip students with foundational AI knowledge that is built in concert with knowledge of responsible AI principles and practices. The availability of digital tools specifically designed to support integrated, experiential learning is limited for now. Future efforts can continue to build off the instructional approaches and models described in this special track of the EAAI 2022 proceedings, leveraging and contributing to the ever-growing availability of software tools to support AI youth learning, such as Teachable Machine (Carney et al. 2020), TensorFlow (Abadi 2016), and AI extensions for Scratch and Snap *!* (Alturayeif, Alturaief, and Alhathloul 2020; Kahn et al. 2018; Druga 2018).

While TechHive AI shows initial promise, additional research is needed to systematically examine the efficacy of this and similar instructional models, to gather evidence for particular instructional strategies or resources to advance sociotechnical learning, and of the feasibility of various models for implementation. Importantly, given historical barriers to access and blind spots in the development and deployment of AI systems, we see it as paramount to center equity in the design of AI-involving learning experiences and in the questions that drive research of those experiences. We see this as critical not only to respond to the expanding need for an AI-literate workforce, but to better position that workforce for innovations that adhere to principles of responsible AI.

Acknowledgments

This research was supported by a grant from the National Science Foundation. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Abadi, M. 2016. TensorFlow: Learning Functions at Scale. In *Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming*, 1–1.

Ali, M.; Sapiezynski, P.; Bogen, M.; Korolova, A.; Mislove, A.; and Rieke, A. 2019. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–30.

Alturayeif, N.; Alturaief, N.; and Alhathloul, Z. 2020. DeepScratch: Scratch Programming Language Extension for Deep Learning Education. *International Journal of Advanced Computer Science and Applications*, 11(7): 642–650.

Amplify Science. 2021. Coherence Flowcharts. https://my.amplify.com/help/en/articles/2999640-coherence-flowcharts.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.

Barton, A. C.; and Tan, E. 2009. Funds of Knowledge and Discourses and Hybrid Space. *Journal of Research in Science Teaching*, 46(1): 50–73.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems*, 29: 4349–4357.

Bransford, J. D.; Brown, A. L.; Cocking, R. R.; et al. 2000. *How People Learn*, volume 11. Washington, DC: National academy press.

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*, 77–91. PMLR.

Carney, M.; Webster, B.; Alvarado, I.; Phillips, K.; Howell, N.; Griffith, J.; Jongejan, J.; Pitaru, A.; and Chen, A. 2020. Teachable machine: Approachable Web-based tool for Exploring Machine Learning Classification. In *Extended abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8.

Chouldechova, A.; and Roth, A. 2018. The Frontiers of Fairness in Machine Learning. *arXiv preprint arXiv:1810.08810*.

Costello, K. 2019. Gartner survey shows 37 percent of organizations have implemented AI in some form. *Gartner*.

Crawford, K.; and Calo, R. 2016. There is a blind spot in AI research. *Nature News*, 538(7625): 311.

Dastin, J. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

Druga, S. 2018. *Growing up with AI: Cognimates: from coding to teaching machines*. Ph.D. thesis, Massachusetts Institute of Technology.

Eubanks, V. 2018. *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Fortus, D.; and Krajcik, J. 2012. Curriculum Coherence and Learning Progressions. In *Second International Handbook of Science Education*, 783–798. Springer.

Garrett, N.; Beard, N.; and Fiesler, C. 2020. More Than "If Time Allows:" The Role of Ethics in AI Education. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 272–278.

Gillen, S.; Jung, C.; Kearns, M.; and Roth, A. 2018. Online Learning with an Unknown Fairness Metric. *arXiv preprint arXiv:1802.06936*.

Greeno, J. G. 1998. The Situativity of Knowing, Learning, and Research. *American Psychologist*, 53(1): 5.

Greenwald, E.; Leitner, M.; and Wang, N. 2021. Learning Artificial Intelligence: Insights into How Youth Encounter and Build Understanding of AI Concepts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17): 15526–15533.

Grosz, B. J.; Grant, D. G.; Vredenburgh, K.; Behrends, J.; Hu, L.; Simmons, A.; and Waldo, J. 2019. Embedded EthiCS: Integrating Ethics Across CS education. *Communications of the ACM*, 62(8): 54–61.

Hajian, S.; Bonchi, F.; and Castillo, C. 2016. Algorithmic Bas: From Discrimination Discovery to Fairness-aware Data Mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2125–2126.

Hampton, L. M. 2021. Black Feminist Musings on Algorithmic Oppression. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 1–11.

Heinze, C. A.; Haase, J.; and Higgins, H. 2010. An Action Research Report from a Multi-year Approach to Teaching Artificial Intelligence at the K-6 Level. In *First AAAI Symposium on Educational Advances in Artificial Intelligence*.

Hmelo-Silver, C. E. 2004. Problem-based Learning: What and How Do Students Learn? *Educational Psychology Review*, 16(3): 235–266.

Huntingford, C.; Jeffers, E. S.; Bonsall, M. B.; Christensen, H. M.; Lees, T.; and Yang, H. 2019. Machine learning and Artificial Intelligence to Aid Climate Change Research and Preparedness. *Environmental Research Letters*, 14(12): 124007.

Jewitt, C.; Kress, G.; Ogborn, J.; and Tsatsarelis, C. 2001. Exploring Learning through Visual, Actional and Linguistic Communication: The Multimodal Environment of a Science Classroom. *Educational Review*, 53(1): 5–18.

Kahn, K. M.; Megasari, R.; Piantari, E.; and Junaeti, E. 2018. AI Programming by Children Using Snap! Block Programming in a Developing Country. *EC-TEL Practitioner Proceedings 2018: 13th European Conference on Technology Enhanced Learning*, 1–18.

Kim, S.; Jang, Y.; Kim, W.; Choi, S.; Jung, H.; Kim, S.; and Kim, H. 2021. Why and What to Teach: AI Curriculum for Elementary School. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17): 15569–15576.

Kolb, D. A. 2014. *Experiential Learning: Experience as the Source of Learning and Development*. FT press.

Lambrecht, A.; and Tucker, C. 2019. Algorithmic Bias? An Empirical Study of Apparent Gender-based Discrimination in the Display of STEM Career Ads. *Management Science*, 65(7): 2966–2981.

Lave, J.; and Wenger, E. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press.

Lee, S.; Mott, B.; Ottenbreit-Leftwich, A.; Scribner, A.; Taylor, S.; Park, K.; Rowe, J.; Glazewski, K.; Hmelo-Silver, C. E.; and Lester, J. 2021. AI-Infused Collaborative Inquiry in Upper Elementary School: A Game-Based Learning Approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17): 15591–15599.

Long, D.; and Magerko, B. 2020. What is AI literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16.

McKendrick, J. 2020. Artificial Intelligence Skills Shortages Re-emerge from Hiatus. *ZD.Net*.

McLennan, S.; Lee, M. M.; Fiske, A.; and Celi, L. A. 2020. AI Ethics is not a Panacea. *The American Journal of Bioethics*, 20(11): 20–22.

Moll, L. C.; Amanti, C.; Neff, D.; and Gonzalez, N. 1992. Funds of Knowledge for Teaching: Using a Qualitative Approach to Connect Homes and Classrooms. *Theory into Practice*, 31(2): 132–141.

Nonnecke, B. 2018. For Better or Worse, Richer or Poorer: The Future of Tech for Good. https://bnonnecke.files.wordpress.com/2018/02/nonnecke_tech-for-good_feb-2018.pdf.

Nonnecke, B.; Sampath, N.; Sistla, M.; and Crittenden, C. 2020. The Future of Public Sector Work: Human-Centered Technology and Policy Strategies. Technical report, https://citrispolicylab.org/wp-content/uploads/2020/09/The-Future-of-Public-Sector-Work-Report-The-CITRIS-Policy-Lab.pdf.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science*, 366(6464): 447–453.

Prince, A. E.; and Schwarcz, D. 2019. Proxy Discrimination in the Age of Artificial Intelligence and Big Data. *Iowa L. Rev.*, 105: 1257.

Ramsey, J. 1993. Developing Conceptual Storylines with the Learning Cycle. *Journal of Elementary Science Education*, 5(2): 1–20.

Reiser, B. J.; Novak, M.; and McGill, T. A. 2017. Coherence from the Students' Perspective: Why the Vision of the Framework for K-12 Science Requires More Than Simply "Combining" Three Dimensions of Science Learning. In *Board on Science Education Workshop "Instructional Materials for the Next Generation Science Standards"*.

Roth, K. J.; Garnier, H. E.; Chen, C.; Lemmens, M.; Schwille, K.; and Wickler, N. I. 2011. Videobased Lesson Analysis: Effective Science PD for Teacher and Student Learning. *Journal of Research in Science Teaching*, 48(2): 117–148.

Saltz, J.; Skirpan, M.; Fiesler, C.; Gorelick, M.; Yeh, T.; Heckman, R.; Dewar, N.; and Beard, N. 2019. Integrating Ethics within Machine Learning Courses. *ACM Transactions on Computing Education (TOCE)*, 19(4): 1–26.

Savery, J. R. 2015. Overview of Problem-based Learning: Definitions and Distinctions. *Essential Readings in Problem-based Learning: Exploring and Extending the Legacy of Howard S. Barrows*, 9(2): 5–15.

Scott, A.; Kapor Klein, F.; McAlear, F.; Martin, A.; and Koshy, S. 2018. The Leaky Tech Pipeline: A Comprehensive Framework for Understanding and Addressing the Lack of Diversity across the Technology Ecosystem. Technical report, https://www.kaporcenter.org/the-leaky-tech-pipeline-a-comprehensive-framework-for-understanding-and-addressing-the-lack-of-diversity-across-the-tech-ecosystem/.

Shaffer, D. W.; and Resnick, M. 1999. "Thick" Authenticity: New Media and Authentic Learning. *Journal of Interactive Learning Research*, 10(2): 195–216.

Tech Can't Fix This. 2020. Technology Can't Fix This. *Nature Machine Intelligence*, 2: 363.

Torp, L.; and Sage, S. 1998. *Problems as Possibilities: Problem-based Learning for K-12 Education*. ASCD.

Touretzky, D.; Gardner-McCune, C.; Martin, F.; and Seehorn, D. 2019. Envisioning AI for K-12: What should every child know about AI? *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 9795–9799.

Verdin, D.; Godwin, A.; and Capobianco, B. 2016. Systematic Review of the Funds of Knowledge Framework in STEM Education. *ASEE Annual Conference & Exposition*, 1–23.

West, S. M.; Whittaker, M.; and Crawford, K. 2019. Discriminating Systems. *AI Now*.

Williams, R.; Kaputsos, S. P.; and Breazeal, C. 2021. Teacher Perspectives on How To Train Your Robot: A Middle School AI and Ethics Curriculum. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17): 15678–15686.

World Economic Forum. 2020. The Future of Jobs Report 2020.

Yu, K.-H.; Beam, A. L.; and Kohane, I. S. 2018. Artificial Intelligence in Healthcare. *Nature Biomedical Engineering*, 2(10): 719–731.

Zwetsloot, R.; Heston, R.; and Arnold, Z. 2019. Strengthening the US AI workforce. *Center for Security and Emerging Technology, Georgetown University*.