Can my algorithm be my opinion?: An AI + Ethics Curriculum for Middle School Students

by

Blakeley H. Payne

B.S., University of South Carolina (2017)

Submitted to the Program of Media Arts and Sciences, School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

© Massachusetts Institute of Technology 2020. All rights reserved.

Author . . . .
Program of Media Arts and Sciences, School of Architecture and Planning
May 28, 2020

Certified by . . . .
Cynthia Breazeal
Associate Professor
Thesis Supervisor

Accepted by . . . .
Tod Machover
Academic Head, Program in Media Arts and Sciences

Can my algorithm be my opinion?: An AI + Ethics Curriculum for Middle School Students

by

Blakeley H. Payne

Submitted to the Program of Media Arts and Sciences, School of Architecture and Planning
on May 28, 2020, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

Children of today can be considered “AI natives.” In the same way that children of the 90s were considered to be digital natives, children of the early 2000s and 2010s have grown up in a world where much of their access to information is mediated by artificial intelligence systems. Furthermore, we expect their futures to be increasingly affected by AI, as consumers and designers.

For this reason, there is a movement to teach AI concepts to K-12 students. Drawing on a tradition of scholarship in Science and Technology Studies and a surge in recent research on the ethical issues associated with the construction of AI systems, it is clear that students not only need a technical education of AI, but an education that will allow them to become conscientious consumers and ethical designers of it.

This thesis presents a set of standards which describe what every child should know about the ethics of artificial intelligence: that it is not an objective or morally neutral source of information and, given that, how to design AI systems with stakeholders in mind. It then describes a series of open-source, largely unplugged activities which address these standards by blending together ethical and technical content. Finally, it presents results from a pilot where students engaged with these activities.

Findings about students’ initial understanding of AI and the ethical dilemmas associated with it are presented, as are students’ understanding after engaging with the curriculum. After participating, students moved from seeing AI as an objective tool to a tool that can be both objective and subjective. By the end of the curriculum, students were able to identify more stakeholders of technical systems and design their own systems according to the values of those stakeholders. This work shows that students can transform into conscientious consumers and ethical designers of AI.

Thesis Supervisor: Cynthia Breazeal
Title: Associate Professor

Can my algorithm be my opinion?: An AI + Ethics Curriculum for Middle School Students

by
Blakeley H. Payne

The following people served as readers for this thesis:

Professor Cynthia Breazeal . . . .
Associate Professor, Program in Media Arts and Sciences
MIT Media Lab

Professor Ethan Zuckerman . . . .
Associate Professor of the Practice in Media Arts and Sciences Director,
Center for Civic Media,
MIT Media Lab

Professor Joseph A. Paradiso . . . .
Alexander W. Dreyfoos Professor,
MIT Media Lab

Acknowledgments

Of the many things I've written throughout my life - papers, blog posts, essays - this acknowledgments section has been the most meaningful, tear-jerking, and joyful to write. To all of you - of whom there are many - infinite thanks. This thesis is not my sole creation, but also yours, in more ways than you will likely ever know.

To the best of husbands, best of men, Mark: I love you. Thank you for all of your support - for quizzing me on SAT vocabulary words in 2011, for asserting that I belong at a place like MIT, for being at my side during the many times I thought I would just quit, and for telling me that it would be okay if I wanted to quit - but also gently reminding me that I am strong. At your request: Bortles!

Mom and Dad: I love you from Cambridge to Columbia and back (and more). There are not enough words to thank you for all that you've given me, but I do want to thank you for instilling in me from a young age an appreciation for education and for letting me ask you endless questions about teaching. I'm really grateful, and extremely proud, to be your daughter.

Matt and Susie: Joining your family has been such a gift. Thank you for the trips up and down the East Coast, for the airport pickups, for reading and editing countless pages about AI and ethics. I feel very blessed to have each of you in my life.

Cynthia: When I was an undergraduate doing research for the first time in an assistive robotics lab, some of the first research papers I read were yours. I remember admiring your work and wondering if someday I could be a researcher like you. Over time, my interests changed (and changed again), and never did I imagine that someday you would become my advisor. But I feel so lucky that all the twists led me here. Thank you so much for giving me a Media Lab home and family. Thank you for listening to my ideas, encouraging them, and becoming an advocate for AI ethics education everywhere. Thank you for supporting me as a researcher, writer, advocate, and whole person. Continue being our fearless leader.

To Abby and Ethan: Thank you both for your time on this thesis but also for making me a better reader, researcher, writer, and a more compassionate person. Not only did I do my best learning and growing in your classrooms, the times I spent learning with each of you are some of the happiest memories of my graduate school career. Thank you.

Joe: Thank you for joining this journey, even in its final hour.

Daniella: Your friendship is the best thing I will take away from my experience at MIT. Your enthusiasm, joy, thoughtfulness, and leadership no doubt improved this work but also made me a better a person, made Personal Robots a better research group, and made the Media Lab a happier, safer, better place to be. I cannot wait to see what great things you will build, and I will always be more than grateful to be your cheerleader, editor, collaborator, and friend.

Kate: There have been many things I have learned since coming to MIT. Some of them have been a joy to learn while others have been a tough lesson - but the most important things I have learned are to be brave in the face of powerful men, and to use my privilege to amplify the voices of others. Both of these things I learned from you. Keep persisting.

Kayla and Maribeth: Y'all are the best friends a girl could ask for, and y'all both do it while being 3,000 miles (or more) away. You are both inspirations to me.

Randi: Thank you for seeing me during my first semester at the Media Lab and helping me to become a happy robot. Thank you for being a fierce advocate for students within MIT and a thoughtful, generous collaborator. Your conviction that kids can learn about and create AI is inspiring, I can't wait to see the many amazing things you will do in the future.

Michael: Graduate school has been no easy task for either of us - thank you for always listening, for solidarity, and math puns that I only sometimes understand.

Judy: You are one of the strongest people I know, and I admire your strength so much because its source is kindness. Thank you for your unyielding support through the many twists and turns we faced together, for listening to me, crying with me, and laughing with me when I needed it the most. You are a brilliant researcher and friend, and I am so blessed to have you in my life.

Zivvy: You have been a great collaborator and friend since day one. So much of what is in this thesis I learned alongside or with you. Thank you for your everlasting enthusiasm, chats about AI, pushing me to be an artsier, weirder, more confident, better version of myself.

To Judy, Zivvy, Abhi, Matt, Isabella, Morgan, Eddy, Richard, Neil, and Pinar: I could not ask for better lab-mates and friends during my first year of graduate school. Thanks for all the thoughtful research, fun conversations, and for being advocates for myself and each other.

To my Personal Robots family: thank you for taking me in, making me feel welcome, and supporting me as a researcher and as a person.

Ilana Pelzman-Kern: Thank you for your help with the summer workshop, for sharing your ideas, and tons of laughter. Keep using your strong, smart, and beautiful voice for good.

Rachael: Thank you for being a beautiful soul and friend, for making it normal to talk about and to chase dreams, for writing poetry, for your introspection, for your ability to listen, and for trailblazing a path to personal fulfillment those of us behind you can follow.

Thank you to Monica and Linda for your patience, guidance, and support.

Thank you to Benjamin Moseley, Roman Garnett, and Gustavo Malkomes for encouraging me to apply to graduate school.

Thank you to Casey Fiesler for your incredible research which paved the way for this work to be done, but also for your kindness and your mentorship.

Thank you to Justin Aglio, Maeve Ronan, and Danny Plouffe for your help with teaching and coordinating the workshops which made this work possible.

Thanks to Alex, Barron, Irene, and Kyle for being so enthusiastic about this work and for bringing Teachable Machines into the world.

Thank you to the many people who took the time to answer my questions about how to improve this work, including Nick Seaver, Alexis Hope, Saber Khan, Jaleesa Trapp, Kade Crockford, Joy Buolamwini, Cathy O'Neil, Kevin Robinson, Sydney Levine, and Jeff Behrends.

Finally, thank you to my University of South Carolina family for challenging me, nurturing me, encouraging me, and lifting me up. Specifically, thank you to Duncan Buell, Jason O'Kane, Jenay Beer, Joshua Cooper, Ed Munn Sanchez, and the many other faculty who encouraged me. I think of you all often. Thank you to Jan Smoak for your unending support and confidence. Thank you to the Land Family for your generous gift in supporting my undergraduate pursuits - you gave a girl from Irmo the opportunity of a lifetime. Forever to thee.

Contents

List of Figures

List of Tables

"Artificial intelligence is tech but also an idea. We're creating ourselves in a machine."

Tess Posner

1 A Note on Terminology

Both of the terms "artificial intelligence" and "ethics" are loaded terms in today's society. Both terms can be used as general shorthand, abbreviations for longer, more holistic meanings, or can be used in very precise ways with quite technical definitions. This can make usage of these terms quite confusing. In this section, I would like to provide some clarity on the usage of these terms in this thesis.

1.1 What Is Artificial Intelligence?

This is a difficult question to answer. AAAI, the largest professional society for those working in the field of artificial intelligence, defines artificial intelligence as "the scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines"[5]. There are two key phrases in this definition which illustrate why this question is so difficult to answer: “mechanisms” and “intelligent behavior.”

When the field was conceived, AI professionals worked on logic and symbolic reasoning, which to today’s AI professionals might seem closer to “pure mathematics” than to the popular or common AI methods of today. The field then moved to expert systems where programs would “derive knowledge” from knowledge bases of facts. Intelligent agents became the next popular focus for AI professionals, where AI systems were imagined as individual agents with goals and means of maximizing those goals. Today, machine learning (often conflated with similar but not equivalent terms like: “data science,” “deep learning,” “neural networks”) is a popular AI methodology [8]. When differentiating machine learning from artificial intelligence, AAAI offers this quote from “The Discipline of Machine Learning” by Tom Mitchell: “The field of Machine Learning seeks to answer these questions: How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?” [17]. There is tremendous debate within the professional and academic AI communities whether any of these methodologies can be considered artificial intelligence. There is debate if the rules-based systems of the past, the “if this happens, then do this”code, should still be considered artificial intelligence in comparison to the so-complicated-that-they-are-often-inexplicable machine methodologies of today. There is debate whether machine learning can be considered artificial intelligence, given its roots in fields like statistics rather than symbolic reasoning. I might point out that it must be difficult to argue if a thing, such as machine learning, belongs in a category as some other things, such as knowledge bases or rules-based systems, when the definition of that thing is a question. All this to say, if artificial intelligence is to be defined by its mechanics, there is no consensus as to what those mechanics are.

Furthermore, what is considered “intelligent behavior?” For this, there is even less consensus, as often the measure of intelligence in a behavior is dependent upon the mechanism. A robot which completes a maze by following a pre-programmed route may appear to be intelligently navigating the maze, but might be considered very different in “intelligence” to a robot which perceives through sensors the maze around it and navigates accordingly.

1.2 What Is a Useful Definition of Artificial Intelligence?

Instead of debating the true nature or best definition of artificial intelligence - should such things exist - I instead ask: what is a useful definition of AI for this thesis? Or, what is a useful definition of AI when educating K-12 students? As will be elaborated on later, I argue that this definition should depict a technology children know in their everyday lives and should empower them relative to that technology.

For this reason, “artificial intelligence” in this thesis is used most commonly to refer to systems reliant on machine learning. This is because the curriculum presented in this thesis heavily leverages currently popular technologies such as YouTube, Google Search, or Instagram, which employ machine learning in their most familiar features. The definition given to children in the curriculum (further described in chapter 7) is that artificial intelligence is any system that takes in a dataset as input, uses a learning algorithm to recognize patterns within that dataset, and outputs a prediction. This definition is not used to exclude other, legitimate types of AI; rather, it is used because the most influential systems in their lives follow this pattern and the ability to recognize those systems as utilizing datasets and making predictions (which is notably different from “knowing an answer”) is empowering.

There is, however, one additional confusion that a definition so inclusive of applications presents which is that the real world implementations of specific, recognized AI methodologies are often bundled together with algorithms not classified as artificially intelligent systems. For example, YouTube’s recommendation system, which suggests new videos for the user to watch, employs categorical AI techniques such as collaborative filtering. However there is a disconnect - not all code that contributes to these recommendations are implementations of AI systems. The code which scrapes tags, likes and dislikes, or comments from videos is categorically not AI. The code which places new recommended videos within the interface is categorically not AI. The code which maintains a user's viewing history is categorically not AI. However all of these pieces together make the AI component effective and useful, and all of them are changeable and susceptible to critique. As Seaver notes, these different pieces of code are often - intentionally and not - referred to as "the algorithm" by those in critical fields like ethics, science and technology studies, and anthropology [88].

1.3 The Algorithm

In this thesis, the terms "artificial intelligence system" and "algorithm" are also used almost entirely interchangeably unless specifically noted. This is in part due to these AI and non-AI bundles like the YouTube recommender algorithm. However, as Seaver notes, there is some risk in using the terms interchangeably. He writes, "Moreover, the 'correct' definition of algorithms has been used precisely to isolate them from the concerns of social scientists and humanists, and it has been picked up by advocates and critics alike to set algorithmic processes apart from cultural ones." That is to say, there is a risk in using the general term "algorithm" that technical people will shrug off valid criticism (and therefore, better ethical practices) using the ambiguity of "algorithm" as a shield for their more specific work (even if the ambiguity is correct for the systems in question). However, Seaver argues that the term "algorithm" is actually useful because it points to the existence of a larger, broader socio-technical system. He writes,

> Following Laura Devendorf and Elizabeth Goodman, I find this a useful way to approach algorithms – not as stable objects interacted with from many perspectives, but as the manifold consequences of a variety of human practices.... If we understand algorithms as enacted by the practices used to engage with them, then the stakes of our own methods change. We are not remote observers, but rather active enactors, producing algorithms as particular kinds of objects through our research. Where a

> computer scientist might enact algorithms as abstract procedures through mathematical analysis, an anthropologist might use ethnographic methods to enact them as rangy sociotechnical systems constituted by human practices. Presuming that algorithms must have a stable and coherent existence makes it harder, not easier, to grapple with their production and ongoing maintenance. [88]

This idea is further supported by Adam Burke in his paper "Occluded Algorithms," where he argues the more ambiguous definition, which is common among non-technical and technical people, is useful when thinking critically about technology. He writes,

> Corporations, states, journalists, technologists, and users of all kinds may also use the term 'algorithm' less precisely to refer to phenomena around computational systems. This important observation can be explained by a secondary, popular definition of algorithm, when viewing an opaque computational system as a disempowered user... Given this, the precision of the secondary, algorithm as system, definition, and its characterization as a popular definition, is preferred. [37]

The purpose in defining AI as a technology which utilizes a dataset to make a prediction, to align that definition with popular, commercial technologies, and the choice to refer to these blurry, applied mixes of AI and non-AI technologies as "the algorithm" is to ultimately point toward the broader social system where these technical systems exist. If a student is able to recognize that a technical artifact they use is a form of artificial intelligence, and that the system uses a dataset and makes a prediction, they are one step closer to considering who is affected by that dataset and how that prediction may help or harm. The use of the term "the algorithm" extends this line of thinking beyond the individual AI system to include to the entire socio-technical ecosystem.

1.4 The Ethics of Artificial Intelligence

Perhaps even more hotly contested than the phrase “artificial intelligence,” is the phrase “ethics,” especially in the context of technology. A person who studies the ethics of AI could study a wide variety of topics: law and regulations to a specific technology or technology companies, the macro-economic or social impacts of a particular technology, the power dynamics within a technology company, constructing particular optimization constraints for an AI system in a very specific domain... and so much more. In 2019, the Berkman Klein Center for Internet and Society analyzed 32 “principle documents” from governments, technology companies, advocacy groups, and more, which outlined each organization’s guidelines and principles around developing ethical and human-rights-centered AI. Their analysis recognized eight large themes in these documents: accountability, fairness and non-discrimination, human control of technology, privacy, professional responsibility, promotion of human values, safety and security, and transparency and explainability [64]. This means that “ethics” is often used as a shorthand for this long list of possibilities. Which topics are covered in this curriculum and why are discussed in later chapters.

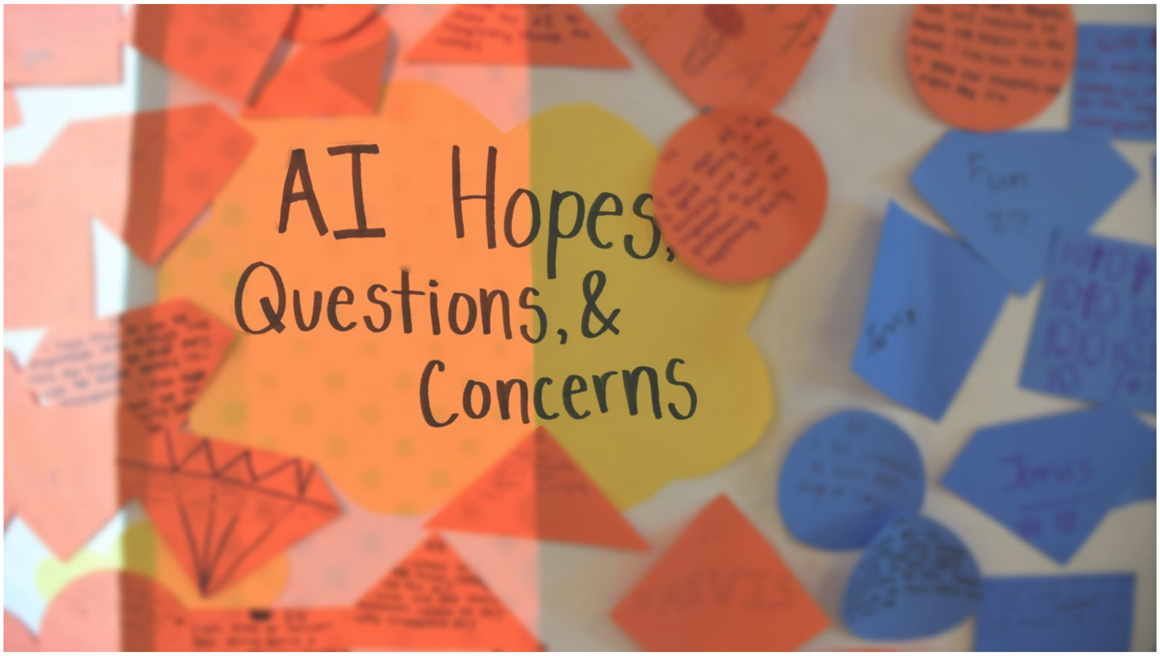


Figure 1-1: A snapshot of a student-composed mural. Students explore what the "ethics of artificial intelligence" means by writing about their hopes, questions, and concerns around AI.

"Algorithms are opinions embedded in code."

Cathy O'Neil

2

Preface

In fall 2016, I was a junior at the University of South Carolina studying computer science. My best friend, Maribeth, was a year ahead of me and had just begun her senior year. And with the arrival of her senior year came the task of finding a "real" post-graduation job.

I remember one Friday night we hung out at her apartment and watched *The Martian*, a movie Maribeth often referenced (some joke about a space pirate or something) but I had never seen. As I settled in to watch the movie, Maribeth opened her laptop and started filling out job applications: typing in her name, her work history, and uploading her very carefully curated résumé.

One of the companies she applied to that night was Twitter, at 7:48 p.m. We stayed up, watched Mark Watney become a space pirate, and agreed to meet up the following day to work on some homework.

I returned to Maribeth's apartment the following day. Her first words to me were: "Twitter has already rejected me."

Less than 24 hours later, on a Saturday, Maribeth had been notified that she wasn't a "good fit" for the role. She was devastated. "Am I out of my league applying to a company like Twitter?" she wondered. "I must be, for them to reject me so quickly."

But here's the thing – Maribeth was a good student. Better than good, actually, one of the best in her class. Despite switching majors from chemistry to computer science, she had somehow maintained a perfect GPA, was a founding member and president of our Women in Computing organization, and had glowing recommendations from a previous software engineering internship. And she had done her homework when it came to constructing her résumé for these applications: she included keywords that were known to be appealing to recruiters and formatted it into one column to be machine readable.

So why had she been rejected so quickly? We had a hunch. Not many students from the University of South Carolina joined companies like Twitter. We also knew that not many Twitter engineers had the title "President of Women in Computing" on their résumé - at the time, less than 10% of software engineers at the company were women. What if Maribeth had been rejected because a computer algorithm had never seen a successful candidate from the University of South Carolina before? Or because most software engineers who got promoted were never affiliated with an organization like Women in Computing?

This was 2016, and although Cathy O'Neil's now famous book *Weapons of Math Destruction* had been published just shy of a month prior and told stories of biased hiring algorithms, it had not yet become a common headline in the news. At the time, a hunch was all we had.

A few weeks later Maribeth and I ventured to Houston, Texas, to attend the Grace Hopper Celebration of Women in Computing. This is where Maribeth would find a job. She was sure of it.

The career fair hall was massive - the size of five football fields. It looked like an IKEA showroom: each tech company had clearly hired an architect and interior decorator to build booths with interactive murals, full service bars, cotton candy machines. It was overwhelming, but immediately Maribeth turned to me and said, "I'm just going to go talk to someone at Twitter. Maybe they won't reject me so quickly in person."

And she was right. Her courage paid off. Later that day, Maribeth started the first round of interviews with Twitter. And several weeks later, she accepted a job at a company that was just the right fit for her: Google.

Maribeth was powerful in this situation, and there were three advantages she had that most people do not. First, she was aware that an algorithm had made a decision about her. Second, she knew this decision was wrong. Third, she ultimately decided that her opinion of herself was more important than the opinion of this algorithm. And she acted on it.

I'd like to underscore that these are hard things to do.

Even when algorithms are recognized, they are often viewed as objective arbiters of the truth – not as opinions. Take for example: the following tweet from a conservative columnist, "Socialist Rep. Alexandria Ocasio-Cortez (D-NY) claims that algorithms, which are driven by math, are racist." Although the tweet boasts hundreds of replies explaining how algorithms can, in fact, be racist, it also boasts 9.8K likes [85]. Mathematics, it would seem, is immune to racism.

Algorithms, and artificial intelligence by extension, are marketed to the public as omniscient and therefore trustworthy. In her book *Hello, World*, mathematician Hannah Fry tells the story of Robert Jones, who was driving through West Yorkshire when he realized his gas tank was almost empty. His GPS, however, offered him a shortcut through a nearby valley to get to the closest gas station. Grateful for modern technology, Robert followed the path, and kept following even after the path turned from pavement to dirt. It wasn't until his car was just about to tip over the edge of the cliff that Robert abandoned his car. The GPS had led him astray. Despite seeing a cliff ahead, and despite driving thousands of miles a week for his job, Robert continued to trust the GPS over his own experience. "I had no reason not to trust it," he later recalled [57].

Moreover, it's not just adults who place authority in these systems. In 2017, two colleagues of mine conducted a study where they had children ages 3-10 play with smart devices such as Amazon Alexa or Google Home. After the children had the opportunity to play with these devices, they were asked a question: do you think this device is less smart, just as smart, or smarter than you?

Among the three and four year olds, 20% of them responded and said the Alexa device was smarter than them. None of the three and four year olds thought Google Home was smarter than them.

However, the six through ten year olds had quite different responses: 100% of them thought Alexa was smarter than them, and 43% reported thinking Google Home was smarter than them [48]. This makes me wonder: how do we ensure that these children have every advantage Maribeth had? How do we ensure these children don't grow up to drive their vehicles off a cliff?

How do we get them to see algorithms as opinions?

"...one of many steps that need to be taken toward this change is at the level of education - and whatever else we can do to ensure that no one is 'just an engineer' anymore."

Casey Fiesler

3

Educating Children in the Era of AI

In February 2018, the Wall Street Journal published an article titled "How YouTube Drives People to the Internet's Darkest Corners," which recounted an investigation the journal had done into YouTube's recommender system. The recommender system, it seemed, suggested increasingly conspiratorial or extreme content to its users, regardless of whether the user was searching for it or not. The article shows screenshots of "suggested videos" that were recommended to a brand-new user with no previous viewing history. The suggestions included videos containing conspiracy theories around the Pope ("How Dangerous is the Pope?" read the top recommendation) and videos implying that the Earth is flat ("Lunar Eclipse doesn't work on your Globe!" read another recommendation) [77].

Just a month later, professor and techno-sociologist Zeynep Tufecki penned an opinion piece in the New York Times called "YouTube, the Great Radicalizer," corroborating the experience described in the Wall Street Journal article. “What we are witnessing is the computational exploitation of a natural human desire: to look ‘behind the curtain,’ to dig deeper into something that engages us,” Tufecki writes [102]. Since the publishing of her op-ed, the New York Times has published several stories portraying how YouTube’s recommendation algorithm has influenced young people’s ideology. They tell the story of Caleb Cain, a man who, after dropping out of college, found comfort in watching self-help videos on YouTube. Those videos recommendations, however, eventually turned into recommendations for white nationalist propaganda. Cain was hooked [84]. These stories also portray the political situation in Brazil, where YouTube’s recommendation algorithm popularized far-right individuals, especially with young men. One of these individuals was Jair Bolsonaro, the current president of Brazil who is known to support human rights violations [55]. Tufecki continues to write:

> In effect, YouTube has created a restaurant that serves us increasingly sugary, fatty foods, loading up our plates as soon as we are finished with the last meal.... This situation is especially dangerous given how many people – especially young people – turn to YouTube for information. [102]

And she’s right.

Pew Research Center reports that 44% of teens are online “almost constantly,” and that 32% of teens use YouTube most often when they’re online. However, it’s not just teens who are on YouTube. Pew also reports that 81% of all parents with children under the age of 11 let their children watch YouTube videos, and it’s known that the same recommender system that prioritizes increasingly conspiratorial and longer content accounts for 70% of watched content on the platform [26].

Children’s interactions with AI do not end with YouTube, though. The children of today live in a world where almost all information and interactions are mediated by artificial intelligence, whether that be getting information from Google search results or chatting with a friend using predictive texting. Common Sense Media reports that 39% of children (the plurality) use social media as their main news source while Pew

Research Center reports that teens are more likely to spend time with their friends online than in person. It's estimated that 16% of Americans have a smart speaker in their home, and many news articles have dubbed these smart speakers as this era's "Mary Poppins." Today's children are not just digital natives, they are AI natives. They are comfortable navigating AI-mediated life. To them, it's normal.

3.1 The Need for K-12 Education

There is a problem with this normality, however. While children - and the public at large - are becoming increasingly dependent on AI-mediated technologies to learn new information or connect with others, AI itself has become more complicated, less transparent, and sometimes invisible to the people who use it. For example, a study in 2015 has shown that 62.5% of Facebook users are unaware that the content on their News Feed is curated by an AI system rather than showing posts in chronological order [50]. Additionally, research has shown that children are willing to disclose personal information to robots and smart toys and are unaware that this information is being recorded by the device, and that parents are extremely worried about their children's trust in AI-mediated toys and devices [105].

Furthermore, although AI is ubiquitous in the developed world, only a small minority (approximately 10%) of Internet users consider themselves to be an expert in AI [67]. According to a report by Tencent in 2017, only 300,000 people in the world have the skills needed to build AI systems [103]. Compare this number to the World Economic Forum's prediction that there will be 58 million new jobs related to AI by 2022 [99].

This number alone motivates the need for widespread AI K-12 education. Many arguments for AI education follow the same logic as the arguments for greater K-12 STEM education efforts: this is where the secure, well-paying jobs will be in the future, so let's get kids interested in the topic from a young age! This sentiment is echoed by the AI4K12 initiative, an organization dedicated to democratizing K-12 AI education in the United States, in their "blue sky" paper [101].

Moreover, this is not the only reason why AI K-12 education is needed. As researcher Randi Williams addresses in her master's thesis, the field of AI is suffering from a "diversity crisis." Although AI is used by many, it is built by few, and a homogenous few at that. Williams writes, "Thousands of people cannot build technology that equitably addresses the concerns of billions" [104]. Many examples throughout history prove Williams correct and show the dangers of having one kind of person (almost exclusively Caucasian men) build technology. Seat belts are known to protect men better than women in car accidents due to crash testing with almost exclusively male dummies [33]. Now, in the era of AI, we know hiring algorithms are biased against women's résumés and facial recognition systems aren't trained to detect darker skinned faces [59, 36]. Williams writes, "How do we start changing the face of AI? One way is by empowering the youngest members of society, the next generation of technologists" [104].

Figure 3-1: A group of 5th, 6th, and 7th grade "AI natives" are excited to learn not only about AI, but also the ethics associated with it.

3.2 The Need for AI Ethics Education

Teaching students how AI works is not enough. In the 1988 text *Education for Democracy*, Jean Piaget wrote,

> The principal goal of education in the schools should be creating men and women who are capable of doing new things, not simply repeating what other generations have done; men and women who are creative, inventive and discoverers, who can be critical and verify, and not accept, everything they are offered.

Indeed, there are many mistakes we do not want the next generation of AI engineers to repeat, from mistakes as simple as building AI that is a bad sport at Tetris to building AI that associates the search term "black girls" with pornography, or that violates laws like the Fair Housing Act [32, 78, 100].

When YouTube employees were designing their recommendation algorithm to suggest videos to users, they were likely trying to answer the question, "how do I build a recommender that suggests videos users would like to see?" rather than the question "how might this "recommender harm communities?" or the even bigger question "should I build a recommender system that measures success in number of views?" It is this line of thinking - or "unthinking" as feminist data scientist Caroline Criado Perez likes to say - coupled with a culture of "move fast and break things" and techno-solutionism that gave us a platform dubbed "The Great Radicalizer" [82].

The next generation of AI designers needs to be prepared to both think of and answer all of these questions. Until this point in time, most AI education has focused on the first question - the so-called "technical question" - and not the latter two, the questions of ethics. It's time for the boundaries to fall.

3.3 Children as Conscientious Consumers of AI

There are reasons beyond a child's possible employment as a software engineer as to why they still need to learn about the ethics of AI systems.

The first reason is the need for AI literacy and the ability to recognize, as exemplified in the preface, that algorithms are like opinions. Children need to know that when they conduct a search on Google, the results displayed are not ranked by "factual accuracy." Rather, they are optimized in such a way that negotiates the user's perceived interests and Google's need to match users with ads from advertisers. As consumers, it is important that children recognize the mechanism responsible for ranking search results is essentially the same as asking for Google's opinion for the "most relevant" results...and knowing that Google's opinion likely favors its own interest over that of the user. Similar to media literacy, this form of AI literacy might translate into a child clicking to the second or third page of search results.

Furthermore, while a child may never become a software engineer, it is likely that their future careers will involve AI. In October 2017, reports surfaced that Amazon was building a machine learning hiring algorithm, and that the algorithm was biased against women.[1] Amazon was ultimately planning to sell this algorithm to other companies to use [44]. Although many children might not grow up to be the engineers who build such an algorithm, they might grow up to be the employee who purchases the algorithm or who uses it to hire new employees into their company. In both circumstances, knowing that an algorithm could be designed and built to pass over qualified candidates is essential knowledge.

3.4 Children as Democratic Designers of AI

In May 2019, San Francisco became the first city in the United States to ban facial recognition technology. In June 2019, it was banned in Somerville, MA, and then in July it was banned again in Oakland, CA [107]. While there have been many hearings before Congress about the use of technology in the United States, each decision was made by informed, local governments [14]. In the case of the Somerville

[1] Some argue that even though the algorithm was known to rate more men higher than women with equivalent qualifications, it might have been less biased than a human recruiter. While this may be true, I argue that (1) without ethics education, personnel involved might never think to test this due to a culture of techno-solutionism, and (2) engineers with ethics training would have been able to seek out even better data to train the system on.

ban, comments from the Security Industry Association, the American Civil Liberties Union, and 119 Somerville residents contributed to this decision [6]. In order to participate in a democratic process such as this, it is important that children as citizens are aware of the potential benefits and harms AI technologies pose.

3.5 Contributions of Thesis

With this motivation in mind, the main contribution of this work is an open-source, largely unplugged AI and ethics curriculum designed for middle school students. This curriculum is unique not only in that it is one of (if not the only) few K-12 AI curricula to have an integrated approach to ethics, it is one of few AI curricula at any level to approach AI from an ethical design standpoint. This curriculum, including materials and teacher guides, can be found at url: `https://bit.ly/mit-ai-ethics`. However, in designing - and later assessing - this curriculum, many other contributions were also made.

First, a review of what educational activities or curricula exist with respect to AI and ethics is given in the next chapter. Chapter 5 gives an overview of the broad, and growing, field of "AI and Ethics," and a framework by which we can prioritize and organize topics into future curricula.

Second, in addition to these reviews, the two pilot studies performed (described in Chapters 7 - 11) provide insight as to what children already know about artificial intelligence and the ethical issues associated with it. Furthermore, key themes from many interviews with educators, philosophers, AI experts, policymakers, and more are provided to give context as to what content should be included in such a curriculum as well as how that content should be taught. The findings from these interviews are presented in Chapter 5 and Chapter 6. All of this information can be used to design future curricula either related to ethical design, artificial intelligence, or both.

Third, lessons learned from two pilot studies are presented. The design recommendations given from the initial pilot are described in Chapter 7 and utilized in a second version of the curriculum. The effectiveness of this second curriculum is assessed in Chapter 10, and design recommendations for similar projects in the future are given in Chapter 11.

Fourth, and finally, this thesis presents an initial list of suggested standards for what all children should know about the ethical design of AI. These standards are presented in Chapter 8, and design recommendations for future iterations of them are presented in Chapter 11.

Through this work, I will answer the following research questions:

1. Which concepts in the realm of AI and ethics should be prioritized for such a curriculum? (Chapters 2, 3, 5, and 11)

2. What are children's preconceptions of artificial intelligence and the ethical issues associated with it? (Chapters 4, 7, 10 and 11)

3. What are children capable of learning about key concepts in AI and ethics? And how does this inform a developmentally appropriate curriculum? (Chapters 4, 7, 10, and 11)

4. How can a curriculum be designed so that it is useful, accessible, supports educators, and is effective in student learning? (Chapters 6, 7, 8, and 11)

5. How does learning about the ethical design of artificial intelligence change children's perceptions of artificial intelligence? (Chapters 7, 10, and 11)

6. How does learning about the ethical design of artificial intelligence change their perceptions of themselves as empowered designers? (Chapters 7, 10, and 11)

It is my hope that this work will provide a path for others - educators, parents, technologists, and more - to encourage the children of today to be empowered ethical AI designers of tomorrow.

"...it is important that we democratize AI now. When anyone can learn about and use AI in creative ways, then AI can become a tool for positive change."

Randi Williams

4

Background and Prior Work

The previous chapter established why we need to educate children about the topic of ethics and artificial intelligence; this chapter reviews what children already know about artificial intelligence and what tools currently exist to educate children on the topic of ethics as it relates to technology and more specifically, to artificial intelligence.

4.1 Children's Knowledge of AI + Ethics

In designing a curriculum on any topic, it is important to ask the following questions:

1. What do the students already know prior to an intervention?
2. What potential barriers to learning can we anticipate and mitigate?

Specifically for this curriculum, it is important to review what children already know about about AI: where it's used, how it works, and whom it affects. It is also important to know what types of AI children interact with in their everyday life in order to best motivate the curriculum to them. This information, in combination with any anticipated barriers students may face in learning about AI, will assist in designing lessons that meaningfully connect with children.

4.1.1 Children's Prior Knowledge

First, it is important to examine research related to what children already know about artificial intelligence and, if possible, the ethical issues associated with it. Although this question is of increasing importance, it still remains largely unclear what children think of artificial intelligence devices.

The large majority of previous research on children's perceptions of AI focuses on children's perceptions of embodied devices such as robots, smart speakers, or smart toys. For example, many works, such as [31], explore how children perceive embodied AI devices to be "alive," while [89] shows children nine years of age or younger are more likely to prescribe particular mental characteristics to embodied AI agents.

These studies have motivated further research to understand what attributes - such as objectivity, authority, trustworthiness, helpfulness, etc. - children ascribe to these embodied devices. For example, [74] shows children find smart toys to be trustworthy, as children would divulge personal information to smart toys, while [105] shows some children (and parents) perceive these devices to be smarter than they are.

At this point in time, there is little parallel research which asks similar questions on disembodied AI agents, such as online recommender or classification systems. While results from [31, 72, 104] suggest that children who have the opportunity to learn technically how these agents work perceive these technologies differently, it is unclear if those findings generalize to disembodied forms of AI.

The best available information about children's prior knowledge of AI systems is about their usage. For example, Common Sense Media reports that children's most preferred news source is social media, with YouTube and Facebook being the most popular platforms among tweens and teens [18]. Pew Research suggests slightly different usage, with YouTube being the most popular social media platform followed by Instagram and Snapchat [28]. While there is little information about students knowledge of how AI works or what their perceptions of AI are in general, we do know which AI systems children interact with and can connect with as examples in the classroom.

Similarly, there is no research at this point in time about children's perceptions of the ethical issues surrounding artificial intelligence - whether they perceive the technology as a whole to be a force for good or bad in society, or what hopes and concerns they might have for the future of the technology. In fact, there is little research on public adult perception of the ethics of artificial intelligence beyond [40]. This thesis aims to close some of these gaps and provide clarity with regards to children's perceptions of AI (beyond embodied agents) and their perceptions of the ethical ramifications of AI.

4.1.2 Anticipating Barriers

In addition to considering what children might already know about AI and ethics, it is important to anticipate what, if any, potential barriers might prevent students from learning about the subject.

One such potential barrier that may exist is children's ability to reason about various ethical dilemmas. As children grow, it is known that their moral reasoning abilities increase. Moral reasoning is the ability to reason about why something might be right or wrong, or the process by which someone determines an action to be right or wrong. The fact that moral reasoning ability changes with age was shown in 1958 by moral psychologist Lawrence Kohlberg, who classified moral reasoning abilities into stages based on his participants' ability to justify particular actions in hypothetical ethical dilemmas [69].

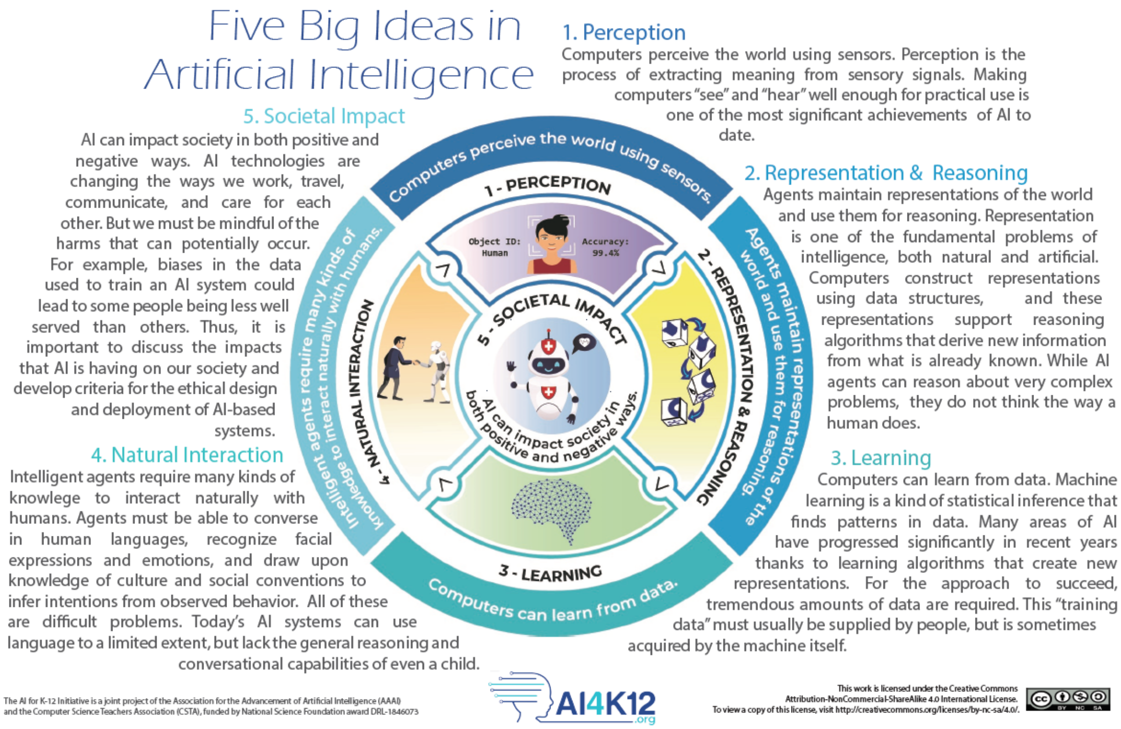
For example, the first stage of moral development is called the "preconventional stage," and children in this stage are likely to choose an action based on their perception of whether the action will be punished or rewarded. Children in this stage are not able to differentiate between a morally wrong action and an action that is punished.

Typically, children move on from this stage of moral development around age ten, or approximately in the fifth grade, to the "conventional stage," where morally good actions are associated with maintaining friendships or social order.

This theory suggests that younger children may struggle with some ethical dilemmas, while older students may be more equipped to inspect morally ambiguous situations.

4.2 The State of AI + Ethics Education

Beyond children's preexisting schema and the potential barriers to their learning, it is important to see what other curricula, tools, and platforms exist which teach artificial intelligence, and in particular, what kind of ethics topics are included and how they are presented.


Five Big Ideas in Artificial Intelligence
1 - PERCEPTION
Computers perceive the world using sensors.
Object ID: Human
Accuracy: 99.4%
2 - REPRESENTATION & REASONING
Agents maintain representations of the world and use them for reasoning.
3 - LEARNING
Computers can learn from data.
4 - NATURAL INTERACTION
Intelligent agents require many kinds of knowledge to interact naturally with humans.
5 - SOCIETAL IMPACT
AI can impact society in both positive and negative ways.
AI4K12.org

1. Perception

Computers perceive the world using sensors. Perception is the process of extracting meaning from sensory signals. Making computers "see" and "hear" well enough for practical use is one of the most significant achievements of AI to date.

2. Representation & Reasoning

Agents maintain representations of the world and use them for reasoning. Representation is one of the fundamental problems of intelligence, both natural and artificial. Computers construct representations using data structures, and these representations support reasoning algorithms that derive new information from what is already known. While AI agents can reason about very complex problems, they do not think the way a human does.

3. Learning

Computers can learn from data. Machine learning is a kind of statistical inference that finds patterns in data. Many areas of AI have progressed significantly in recent years thanks to learning algorithms that create new representations. For the approach to succeed, tremendous amounts of data are required. This "training data" must usually be supplied by people, but is sometimes acquired by the machine itself.

4. Natural Interaction

Intelligent agents require many kinds of knowlege to interact naturally with humans. Agents must be able to converse in human languages, recognize facial expressions and emotions, and draw upon knowledge of culture and social conventions to infer intentions from observed behavior. All of these are difficult problems. Today's AI systems can use language to a limited extent, but lack the general reasoning and conversational capabilities of even a child.

5. Societal Impact

AI can impact society in both positive and negative ways. AI technologies are changing the ways we work, travel, communicate, and care for each other. But we must be mindful of the harms that can potentially occur. For example, biases in the data used to train an AI system could lead to some people being less well served than others. Thus, it is important to discuss the impacts that AI is having on our society and develop criteria for the ethical design and deployment of AI-based systems.

The AI for K-12 Initiative is a joint project of the Association for the Advancement of Artificial Intelligence (AAAI) and the Computer Science Teachers Association (CSTA), funded by National Science Foundation award DRL-1846073

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/.

Figure 4-1: The "Five Big Ideas in Artificial Intelligence," as identified by the organization AI4K12.

4.2.1 K-12 AI + Ethics Education

At the beginning of 2019, the organization AI4K12 published a paper: "Envisioning AI for K-12: What should every child know about AI?" [101]. AI4K12 is an organization jointly sponsored by the Association for the Advancement of Artificial Intelligence (AAAI) and the Computer Science Teachers Association (CSTA) whose goal is to both develop national guidelines for K-12 AI education and to develop a repository of resources to enable teachers to bring AI to the classroom. The paper they published was a "blue sky" paper - it outlined the current state of AI education tools as well as what future tools should be like. Part of the paper included a list of the five "big ideas" in AI, as seen in Figure 4-1:

1. Computers perceive the world using sensors.
2. Agents maintain models or representations of the world and use them for reasoning.
3. Computers can learn from data.
4. Making agents interact comfortably with humans is a substantial challenge for AI developers.[1]
5. **AI applications can impact society in both positive and negative ways.**

Since the paper's publication, the number of tools and resources for K-12 AI education has been rapidly expanding. However, despite "societal impact" being listed as one of AI4K12's five "big ideas" for K-12 AI education, there are still currently very few tools available to educators which incorporate a discussion of societal impact or ethics into AI education. Below I discuss the many categories of resources available to educators and which tools include some form of ethics.

Extensions

Many tools, such as Cognimates, eCraft2Learn, and Machine Learning for Kids extend the capabilities of popular visual programming languages such as Scratch or App

[1] This language has since been changed to "Intelligent agents require many types of knowledge to interact naturally with humans." [45]

Inventor and allow children to independently build, train, and test their own AI systems [10, 11, 16]. Many of these platforms hook into popular AI APIs, such as IBM Watson, to allow students to build on top of preexisting project templates. Of these tools, eCraft2Learn is the only platform which calls attention to an ethics issue by including a video about algorithmic bias as a part of a project template. However, the project template only focuses on awareness of the issue and does not provide guidance on how to avoid algorithmic bias from a technical or design perspective [66]. This is unsurprising as these platforms function more similarly to a programming language, and they are designed to be used as a teaching tool within a curriculum and intended to be paired with guided instruction as opposed to self-contained teaching modules.

Visual Explainers and Games

Other educational AI web tools function less as construction tools and more as explanatory tools. Examples from this category include tools like the many AI Experiments by Google, TensorFlow Playground, and Google's "What If?" tool [2, 22, 24]. These tools, through games or visualizations, give students an intuition how various kinds of AI (typically neural networks or generative adversarial networks) work. Again, in this category, only one of these tools explicitly delves into an issue of ethics.

Google's "What If?" tool acts as a visual explainer for algorithmic bias and shows the user how an algorithm can have disparate outcomes for different groups as training data is changed or as different loss functions are minimized. This particular tool, however, is likely to be reserved for very advanced high school classrooms, as it was designed to be a tool for AI engineers and requires a large quantity of preexisting, expert-level technical knowledge.

Curricula

Several AI curricula do exist at the K-12 level. Some of these exist primarily as research platforms and show the capability of students at various ages to grasp AI and machine learning concepts or explore the advantages of various pedagogies. One such example includes the PopBots curriculum, a toolkit for children ages 4-6 years old to

#	Platform	Topic	Type	Age Range	Cost	Ethics
1	Calypso for Cozmo [9]	Rules-based systems, Perception, State machines	Programming language, Lesson plans	8-15 years old	Requires Cozmo Robot; $14.99	
2	ECS Alternate Curriculum on AI [87]	Neural networks, Algorithmic bias	Lesson plans	High school	Free	✓
3	Elements of AI [79]	Bayesian probability, Neural networks	Online modules	Advanced high school	Free	✓
4	ReadyAI Curricula [83]	Varied topics across Big 5	Lesson plans, Teacher training, hardware	5-14 years old	Free - $3,000	✓
5	Techsplorer by NVIDIA and Technovation [19]	Neural networks, Self-driving cars	Lesson plans	9-14 years old	Free	

Table 4.1: Existing K-12 AI curricula.

explore AI topics alongside a programmable social robot [104]. PopBots showed that preschool age children are able to understand concepts like supervised machine learning, portrayed the benefits of constructionist pedagogy, and demonstrated the impact early STEM education can have on children's perceptions of their own scientific and engineering abilities.

Other curricula are currently available to educators, as depicted in Table 4.1. This list is rapidly expanding in real time however, and more resources can be found at aieducation.mit.edu.

It is worth noting that three of the five resources here contain an ethics component, all of which reserve ethics content to the last modules of the curriculum.

All of the ECS curriculum (#2), Elements of AI (#3), and ReadyAI (#4) curricula include a unit or module on the societal impact of artificial intelligence at the end of the curriculum, typically as the final module, and typically on the topic of algorithmic bias or the impact of automation on the workforce. The ECS curriculum (#2) also teaches on the topics of algorithmic bias and about the impact of automation, but also discusses AI's impact on the future of news media.

AI for Good

In addition to these tools and curricula, there is also a current trend toward AI education events, extracurriculars, and summer programming. Many of these experiences fall under the umbrella of "AI for Good." AI for Good is both a United Nations platform in partnership with technology companies, academic institutions, and national governments, as well as a flavor of computer science project found in academic and STEM outreach programs - the latter is the focus of this section. AI for Good projects are typically characterized by the pairing of a computer scientist to a community issue (at the UN level, these issues are typically related to the UN's sustainability goals, while in the classroom these issues are often related to the local community) and having said computer scientist use AI or machine learning methods to "solve" the issue [4].

These experiences include WAICY, or the World AI Competition for Youth, or teen-targeted summer programs like AI4All or Teens in AI [25, 3, 21]. Students begin each of these experiences by participating in an intensive, introductory AI bootcamp. After learning "the basics," students are divided into teams and asked to identify a community issue that could be remedied with an AI solution. In particular, WAICY asks students to focus on solutions that also utilize AI concepts and robotics concepts since the competition heavily leverages a Cozmo robot [25].

Similarly, the Technovation AI Family Challenge guides parents and children through a similar process at home. The Family Challenge begins by introducing technical concepts like neural networks to children and eventually has them apply their learning using one of the aforementioned tools, such as Machine Learning for Kids, to build an "invention" that will assist their local community [20].

The trend toward AI for Good inspired outreach is unsurprising as there is a large amount of evidence which suggests students, especially underrepresented students, are more likely to become interested in a STEM topic if the topic is presented in relationship to social justice issues or service learning [34, 68]. However, there is some debate if AI for Good experiences can be considered ethics training for AI developers [71]. The following chapter discusses the relationship between the notion of "AI for Good" and other perspectives on the ethics of AI.

Digital Citizenship and Media Literacy Curricula

A few efforts exist which aim to teach that the Internet does not necessarily relay neutral, authoritative sources of information. Notably, there has been a trend in recent years of updating "digital citizenship" curricula. Digital citizenship curricula exist to teach children to be critical thinkers as they navigate the web, and tend to have many media literacy learning objectives. Two prominent curricula exist in this space: Common Sense Media's "Digital Citizenship" curriculum as well as Google's "Be Internet Awesome" curriculum. Both curricula focus on teaching the importance of keeping identifying information private, being kind to other Internet users, as well as identifying the bias of information sources. While neither of these curricula seek to give students an understanding of how AI works (either from a technical perspective or a societal one), both of these curricula have been recently updated to make students aware of issues exacerbated by artificial intelligence. For example, both curricula now include lessons on false news, misinformation, and how misinformation can be spread by algorithms. Google's "Be Internet Awesome" curriculum includes lesson plans on how to identify bots and about the fallibility of digital voice assistants' responses [7]. Common Sense Media's curriculum now includes lessons about the existence of filter bubbles and how filter bubbles can affect one's ability to find balanced information [13].

EthicalCS

Another movement which does not seek to specifically educate children on the topic of artificial intelligence but does seek to highlight the subjective nature of algorithms and their impact on society is EthicalCS. EthicalCS is a community of researchers, educators, and computer science professionals who seek to provide resources to teachers so that they may center ethics in their computer science teaching. Resources provided by EthicalCS include a range of unplugged (no computer needed) activities and discussion based activities on topics such as bias in datasets and questioning whom an algorithm serves. Most of the activities provided are constructed as single activities to be integrated into a larger, technical computer science course or as amendments to popular computer science lessons, and are not meant to be concatenated into a curriculum [12].

4.3 AI + Ethics in Higher Education

Although this curriculum focuses on middle school education, it is worth examining AI + ethics at the undergraduate and graduate level. When examining AI+ ethics education in higher education, it becomes unsurprising that the majority of K-12 AI resources leave out an ethics component altogether, or that many resources that do include ethics content, relegate that content as the last content to be learned. These patterns strongly reflect the teaching practices, content focuses, and standards at the undergraduate and graduate level of study.

4.3.1 Toward Integrated Ethics

In 2014, computer science professor Arvind Narayanan and philosophy professor Shannon Vallor wrote an article entitled "Computing Ethics: Why Software Engineering Courses Should Include Ethics Coverage." In the article, they argue that college software engineering courses should include ethics education. They write:

> Habits are powerful: Students should be in the habit of considering how the code they write serves the public good, how it might fail or be misused, who will control it; and their teachers should be in the habit of calling these issues to their attention.... What matters in these exercises is not

> that students can arrive at the “right” answers; nor even that the instructor have them in hand. In many real-life cases there is no single right answer, only a range of more or less ethically informed and wise responses. What matters is that students get comfortable exercising ethical discernment in a professional context alongside their peers. [76]

They argue for an “integrated” approach to ethics education in the computer science classroom. The logic is that it is impossible to form thoughtful habits in a standalone course, and that the habits must be generalized across all content areas: how can students be mindful of ethics in their software engineering courses, in their algorithms course, in a user experience design course, or in an AI course.

There is another danger in siloing ethics topics from technical topics - that students may grow to believe that ethics is something to consider after the technical considerations or that it is a specialization that they are not responsible for considering. As Casey Fiesler writes:

> It suggests that ethics is not a thing that everyone should be thinking about — not, as Zunger said, “the foundations of all design” — but is, instead, only a specialization. I just build things; someone else can think about the ethics. [52]

In fact, there is evidence to suggest that isolating ethics to its own course or module within a course leads students to perceive ethics as unrelated to their studies or as a specialization in addition to their technical work [95, 46].

Although some argue that dispersing ethics content across multiple computer science classes might lead to it getting cut from those classes when instructors are pressed for time, many universities have now moved to embrace this integrated approach to ethics at the undergraduate level [58, 39, 47, 60, 92]. Due to this movement, there are an increasing number of papers which outline different strategies and topics for teaching ethics content to students.

Given this shift toward more ethics education, one might naturally wonder - what is being taught? A recent analysis found that of college machine learning courses, while the majority do not cover any ethics topics at all, those that do cover topics under the following five themes:

1. Accountability and responsibility.
2. Data privacy and anonymity.
3. Data availability and validity.
4. Model and modeler bias.
5. Model transparency and interpretation.

More broadly, an analysis of 115 technology and ethics courses found that the most common topics included law and policy, privacy and surveillance, philosophy, human rights, artificial intelligence, and social and environmental impact. Additionally, this analysis found that the most common stated outcome for these courses was for students to be able to “critique” technology, with only a few courses wanting students to be able to “create solutions,” “consider consequences,” or “apply rules” [53].

However, just because these are the topics being taught at the university level does not mean that these are the topics that should be prioritized at the K-12 level. Answering the question "what *should* we be teaching when we teach ethics?" is the goal of the next chapter.

"AI for Good is easy;
it's AI-That's-Not-Bad
that's hard."

Vivienne Ming

5

What Should We Be Teaching When We Teach "AI + Ethics"?

At the beginning of this thesis, I provide a note on terminology, stating that "AI ethics" is a kind of catchall phrase with no precise meaning. Instead, the phrase is used in a variety of contexts such as in law and regulation, in the context of doing "social good" for communities with AI, in the context of mitigating the potential harm of AI, and many more. The educational AI resources in the previous chapter reflect this: while most projects do not include any ethics teaching objectives, those that do are widely varied in what that ethics objective might be. This leaves us with the question: what should kids know when it comes to "AI ethics"? In this section, I look to the curriculum's motivations in order to prioritize which lessons students should learn first and outline them below.

The main focus of this chapter is this question: *what should we be teaching when we teach AI ethics?* The next chapter focuses on the question of whom is this curriculum for and how it can be designed to serve them best. It is worth noting, however, that these three questions of what should be taught, whom should be taught, and who is teaching, cannot be fundamentally separated. It is the combination of the answers to these two questions that ultimately determine which learning objectives are prioritized for the suggested standards outlined in Chapter 8.

5.1 Popular Notions of AI + Ethics

It is, however, worth taking some effort to acknowledge popular first impressions on the ethics of artificial intelligence and how they relate to this curriculum and others.

5.1.1 The Problem with Killer Robots

One problem with the labeling of "AI + Ethics" is that artificial intelligence can be used to describe very real technical systems that exist in the world today as well as creations of science fiction and fantasy. A common misconception is that "artificial intelligence" actually refers to "artificial general intelligence (AGI)." This term refers to a kind of AI that is just as knowledgeable as humans and can "think for itself." When all kinds of AI are misrepresented with artificial general intelligence (as often occurs in popular film franchises like *Terminator* or *The Avengers*), "AI ethics" is as misrepresented as a field that is only concerned with preventing "killer robots."

This misconception is furthered by prominent men in science and technology such as Stephen Hawking or Elon Musk who espouse their belief about AGI being the greatest existential threat to society. While questions about how AGI should be built and what AGI can teach us about human nature are interesting, narrowing "AI Ethics" to solely mean the study of the impact of this kind of non-existent AI is a distraction from real systems which cause real harm - and therefore, a harmful practice.

Instead, it is more pressing that designers of technology focus on mitigating the harm of "narrow artificial intelligence" (abbreviated NAI), or artificial intelligence which is designed only to perform a specific task. Innocuous examples of NAI include algorithms which classify email as spam (or not) or algorithms which play games of chess. However the narrow AI of today also includes higher-stakes algorithms, everything from predicting if someone can pay back a loan, has an "appropriate" personality for a particular job, or will commit a crime. As Cathy O'Neil writes:

> I look around, I realize there is no need to imagine some hypothetical future of human suffering. We are already here. Data scientists are creating machines they do not fully understand, machines that separates winners from losers for reasons that are already very familiar to us: class, race, age, disability status, quality of education, and other demographic measures. [80]

Issues related to the ethics of NAI are gaining more and more attention as headlines about "sexist," "racist," and "biased" AI systems, not about killer robots or ominously polite computer voices [75, 65, 93]. So now the question becomes: what do our students need to know about the ethical ramifications of today's artificial intelligence?

5.2 Becoming Conscientious Consumers

5.2.1 Seeing Technology as Political

In his 1985 paper, "Do artifacts have politics?", Langdon Winner writes about the especially low-clearance bridges over the parkways on Long Island. These bridges have a clearance of nine feet - which might not sound unusual until you realize that standard overpasses in the United States have a clearance of 16 feet.

When hearing about these bridges, it is easy to think, "Oh, there must be some structural reason why those bridges had to be built with a clearance of nine feet." This was not the case. As Winner reveals, "It turns out, however, that the two hundred or so low-hanging overpasses on Long Island were deliberately designed to achieve a particular social effect." Robert Moses, who held over ten titles related to parks and parkway authority from 1924 to 1975, designed the bridges with a clearance of nine feet because the public transit buses, which primarily served racial minorities and low-income communities, were twelve feet tall. What now appears as “just a bridge” was actually a tool for segregation [106].

Winner’s point in telling this story - as perhaps the title gives away - is that all artifacts, from architectural objects to code, are political. There is no such thing as a morally or politically “neutral object,” a technical detail like the height of a bridge can make society fairer or more unfair and have a lasting impact on a community. Moses intentionally built his racism into these bridges, but it is not uncommon for unconscious bias to also creep into most innocuous designs. For example, when Apple’s HealthKit app launched in 2014, it could track a user’s blood alcohol content - but not user’s period. Why was it launched without a standard health tracking feature? Because those who built it did not have periods to track. This result was not produced from malicious intent. Rather, it was a reflection of the opinions and experiences of its makers [49].

This sentiment, although it is not new in communities which critique technology, is seemingly new to computer scientists. In 2018, Science magazine reported on the development of a new algorithm built to predict if a crime was gang-related or not. The algorithm seemed to perform spectacularly by reducing 30% of the errors the previous model had made, but this was not the focus of the story. Instead, it focused on an answer the algorithm’s engineer gave to a question during an AI conference.

After presenting his work, the engineer was asked how he could ensure the tool wasn’t biased or what should be done if it was found that the tool misclassified a crime as gang-related. The engineer’s response was: “I’m just an engineer” [63]. This response is unsurprising given the lack of emphasis on ethics and societal impact in computer science programs, as evidenced in the previous chapter. It is also a particularly dangerous response. There are two implications associated with the phrase “I’m just an engineer.” The first is that engineers should build whatever they would like to, because they can. The second is that engineers can abdicate all responsibility for the consequences of something they build.

AI is particularly not neutral, and rhetoric that claims it is so makes it even harder to see. In designing this curriculum, I interviewed many experts who are working in various areas of AI ethics. They included a data scientist who is also the father of a middle school student, a previous hedge fund data scientist turned algorithmic auditor, an anthropologist with a focus on recommender systems, a professor of philosophy who teaches a graduate course on AI and ethics, and a director at the American Civil Liberties Union who focuses on protecting civil liberties in the era of AI. Each of them, when asked what the most important takeaway of the curriculum should be, said, "Technology is not neutral." The professor said, "This is my biggest challenge, getting my computer science students to see that they are already making ethical decisions."

5.2.2 From "What Ethics?" to "Whose Ethics?"

A corollary of the idea that all artifacts, especially algorithms, have politics, is that ethics already are, and have always been, a part of AI. This is key insight: as public discussions of the societal consequences of technology become more pervasive and as dedicated conferences like AIES at FAT*, and FAccT [1] increase, it is easy as researchers, educators, or citizens to fall into the trap of thinking that "ethics" is just now becoming integrated into the field of artificial intelligence and machine learning. But this isn't true: someone's ethics have always been implemented in our technologies, in the same way that Moses's ethics were embedded in the bridges he created. This concept becomes clearer when you consider the definition of ethics as the "moral principles that govern a person's behavior or the conducting of an activity."

So the critical question we want our children to be able to think about is not: what is ethical AI? Rather, we want them to be able to ask: whose ethics are being accounted for?

[1] It is worth noting that the key similarity between conferences like AIES, FAT*, FATML, and others is that they are conferences hosted by professional *computing* organizations that focus on the ethics and politics of technical artifacts. Other communities, such as the Science, Technology, and Society community, have been discussing these topics and hosting conferences since the 1960s, and have not been given the same spotlight as the computing community.

5.3 Becoming Ethical and Democratic Designers

In her SXSW keynote titled “You Think You Want Media Literacy... Do You?”, danah boyd interrogates the need for additional media literacy education in K-12. She argues that if done poorly, media literacy education can actually produce more harm than good. Often, media literacy is equated to critical thinking skills that ask students to doubt what they see, read, and watch. She writes:

> But the hole that opens up, that invites people to look for new explanations... that hole can be filled in deeply problematic ways. When we ask students to challenge their sacred cows but don’t give them a new framework through which to make sense of the world, others are often there to do it for us. [43]

The same is true for teaching kids to be critical of technology. If we are to shift children’s mindsets from seeing technology as neutral or objective, what are we going to offer them to help stabilize their worldview without simply telling them what to think or believe?

boyd offers teaching empathy to others as one incomplete solution to this task with the warning, “Empathy is a powerful emotion, one that most educators want to encourage. But when you start to empathize with worldviews that are toxic, it’s very hard to stay grounded.” Building on this idea she suggests teaching kids to understand how different people can construct knowledge:

> From an educational point of view, this means building the capacity to truly hear and embrace someone else’s perspective and teaching people to understand another’s view while also holding their view firm. It’s hard work, an extension of empathy into a practice that is common among ethnographers.... The goal is to understand the multiple ways of making sense of the world and use that to interpret media. [43]

If, as I state above, the question we want students to think critically about is whose ethics are being accounted for, I think boyd’s instruction would ask us to also teach children to think about how we might design technology given such diversity in experiences and worldviews.

5.3.1 The Need for Inclusive Design

It is impossible to answer the question "whose ethics is being accounted for?" and to think about more inclusive design without discussing issues of diversity in the tech field, of those who are actually designing and creating these pervasive technologies. As Mia Dand, the founder of the global Women in AI Ethics initiative, says, "Diversity and AI ethics are not separate issues. You can't have ethics without diversity" [98]. It is well known that the technology industry is overwhelmingly comprised of men, and generally Caucaisan or Asian men. Between Apple, Microsoft, Facebook, and Google, the percentage of women in technology roles as opposed to men ranges from 19.9% at Microsoft to 25.7% at Google. There is no data for non-binary individuals. Amazon does not report gender breakdown for technology roles, only for all roles. The data shows even less diversity when accounting for race and ethnicity. Among the aforementioned companies, none of black, Hispanic, Native American, or "other" groups accounted for more than 8% of technical developers. In general, less diversity is represented in management and leadership roles. The numbers in AI and machine learning are even worse - it is estimated that 12% of contributions at leading machine contributions were made by women, and companies like Facebook estimate 15% of their AI professionals are women [91]. None of the companies provide intersectional data [1, 73, 81, 96].

As previously mentioned, many technologies, even safety critical ones such as seat belts, are often designed for men by men [33]. Given the data in the previous paragraph, it is frustrating yet unsurprising to learn that many major commercial facial recognition systems work better for male faces than female faces or lighter skinned faces than darker skinned faces, and that these systems perform their worst when presented with darker, female faces [36].

It is clear that efforts to make the technology industry a more diverse and inclusive place are essential pieces in building better, fairer, more inclusive AI. However, we must ask: are engineers' ethics the only ethics that matter when building a technology? Given the global reach of technology, especially platforms like Twitter, Facebook, or YouTube which have been attributed to sway democratic elections, the answer seems to be a resounding "no" [55]. Other stakeholders must be accounted for in the design of these far-reaching technologies.

5.3.2 Designing with Stakeholders in Mind

Who are the other stakeholders in an AI (or any socio-technical) system, and how do we incorporate their needs into the AI design process? As human-computer interaction scholar Katie Shilton writes in her paper "Values and Ethics in Human-Computer Interaction":

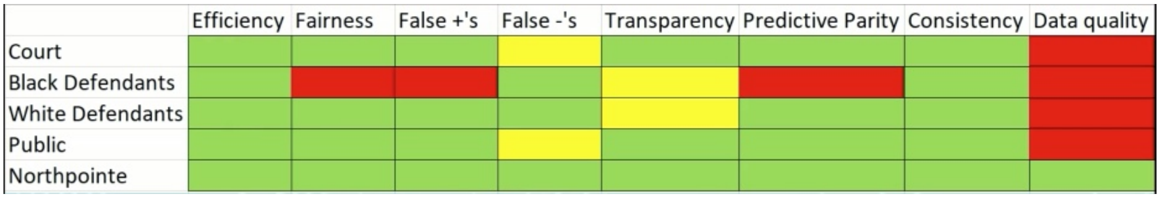
> Direct and indirect stakeholders of technologies are difficult to enumerate. Our design practices may impact people beyond our users, whether through the collection and use of information about people during design, through secondary unintended consequences, or because of the natural resources our technologies use.[90]

Processes like value sensitive design, participatory design, and design justice offer protocols for designers to follow that seek out and respond to the needs of various stakeholders. Shilton writes, "Batya Friedman's work on value sensitive design (VSD) has been some of the most influential in uniting the space of computer ethics and methods for design, and broke new ground for generative approaches to values in design." The VSD process begins with designers identifying a set of principle values which are intended to guide the rest of the design. Commonly cited values include things like privacy or equity. The process then continues iteratively with empirical investigations and design play [56].

Participatory design is in response to critiques of value sensitive design, particularly the critique that any set of values could be universal and that even if such a set existed, that those values could be determined in a top-down approach by the researcher. Instead, participatory design encourages researchers and designers to embed themselves in local communities and designing with, as equals, to those who are being designed for [90]. Design justice is a particular type of participatory design which asks designers to prioritize dismantling the matrix of domination, or the ways in which oppression by race, class, and gender intersect. The primary method for doing so is to prioritize the needs of stakeholders who are at risk for the most harm to be done by the technology. This means prioritizing the voices of people of color, queer voices, Indigenous voices, and many others [42, 41].

Unfortunately, participatory design and design justice are hard to accomplish in the classroom due to the logistical challenges of connecting children to various stakeholders. For this reason, this work leans heavily on VSD because it does a good job of making the political aspects of technology explicit, but it is important to discuss and emphasize with children - as danah boyd does - that empathy can only go so far. The best processes involve talking with and actively listening to various stakeholders, especially those who are potentially affected the most by the technology in question.

The Ethical Matrix



	Efficiency	Fairness	False +'s	False -'s	Transparency	Predictive Parity	Consistency	Data quality
Court	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟥
Black Defendants	🟩	🟥	🟥	🟩	🟨	🟥	🟩	🟥
White Defendants	🟩	🟩	🟩	🟩	🟨	🟩	🟩	🟥
Public	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟥
Northpointe	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

Figure 5-1: An example of an ethical matrix on the problem of recidivism risk scoring algorithms. Reprinted from *Weapons of Math Destruction, Ethical Matrix, Nate Silver and More Highlights from the Data Science Leaders Summit* at https://www.kdnuggets.com/2018/07/domino-data-science-leaders-summit-highlights.html.

One tool to help elucidate embedded values in technical systems is the ethical matrix. Originally a tool used in bioethics translated to the technology setting, Cathy O'Neil recently has written about how it can be used in the context of artificial intelligence [86, 61]. The ethical matrix is a 2-dimensional table where stakeholders are listed on one axis and the values those stakeholders hold in the system are listed on the other axis. An example of an ethical matrix is shown in Figure 5-1. Designers of a new technology can then go row by column and identify where stakeholders' values align and where they conflict. Designers can also identify which conflicts in values might produce the most harm for any of the stakeholders involved. In filling out the matrix, designers are forced to recognize that multiple diverse stakeholders exist within the system, and to empathize with multiple perspectives of a diverse set of stakeholders. For this reason, this tool is heavily leveraged throughout the curriculum.

5.3.3 A Note on AI for Good

There is another popular notion of "AI + Ethics," especially in the education sector, which conflates the field with those who try to leverage AI for the "social good." Part of this sentiment arises from the tech industry itself, where the ethos is that the tech industry exists to make the world a better place. It is likely that the phrase "AI for Good" is patterned after another common project "Technology for Social Good." However, time and time again we have seen well-intentioned technological solutions cause real harm to marginalized communities. As Mark Latonero writes, "While AI for good programs often warrant genuine excitement, they should also invite increased scrutiny. Good intentions are not enough when it comes to deploying AI for those in greatest need" [71]. This is exemplified by many "technology for social good" projects, such as the proposal presented in the article "How Soylent and Oculus Could Fix The Prison System (A Thought Experiment)." In this article, the author, Shane Snow, suggests that all prisoners have their daily meals replaced with Soylent, a nutritional beverage, and placed in isolation with virtual reality headsets in order to maintain "safe" social interaction [94]. The proposal was well-intentioned but received tremendous backlash. Now, above the original post, are amendments. Snow writes:

> Instead, I would have started out by breaking down the fundamentals of prison's problems and the elements of what it takes to take care of and rehabilitate prisoners... and then I would have gone out and included

> a ton of experts in the process of brainstorming different approaches of prison reform based on these first principles. I would not have dashed this off from just my own limited point of view. [94]

Educational AI for Good projects can, unfortunately, fall prey to the same mistake Snow did. For example, in April 2019 a video titled "My Drawings Speak Up" surfaced, where a child who had participated in an AI for Good challenge presented her project: a tool that predicts if a child is being abused based on their classroom drawings. Again, while the tool was well-intentioned and demonstrated a clear desire to meet a real world need with technology, there were many unchecked red flags. For example, the curated training dataset was small, vetted by only one psychologist, and collected in the public school setting [70]. If deployed (or even continued to be developed), the project could have caused a great deal of harm, which could have been originally avoided if stakeholders were consulted.

In the context of education, I think it is important to acknowledge that there is a temptation to demonstrate that AI has societal impact by engaging in "AI for Good" projects, and then equating Big Idea #5 "societal impact" with ethics. It is important to note that discussing the fact that AI has a societal impact is not the same as discussing what that societal impact could or should be.

5.4 Moving from AI Ethicists to Ethical Designers of AI

If students only learn two ideas from this curriculum, I wish that they be the two outlined above:

1. That algorithms, and particularly AI, are not "morally neutral." Rather, they are more like opinions in the way they represent the world.
2. That students would design algorithms, and particularly AI, with a diverse set of stakeholders, and especially the most vulnerable stakeholders, to make the opinions of their algorithms more just.

In the last chapter, I mentioned a study which analyzed tech ethics classes at the collegiate level. While the most commonly covered topics in these classes included topics such as law and policy, privacy and surveillance, philosophy, human rights - design was one of the least mentioned topics. Among the learning outcomes, “see multiple perspectives” and “create solutions” were ranked in the middle, below outcomes such as “spot issues” or “make arguments” but above “consider consequences” and “apply rules” [53]. I agree that outcomes such as “spot issues” or “make arguments” do turn students into more conscientious consumers, able to see the limitations of the often heralded technology in their lives. However, per boyd’s earlier argument, it is not enough. We must not only teach our students to identify “problematic” technology, we must also equip them with tools that make them feel empowered, but also do not grant them power that may harm marginalized communities. We must prepare our children not only to be ethicists, but ethical designers of AI.

"Instead of saying 'deploy' technology, I prefer to use the word 'integrate.' Because it prompts the question 'into what?'"

Madeleine Clare Elish

6

Designing the Curriculum

In addition to what should be taught, who is being taught and who is doing the teaching are fundamental questions to consider. The goal of this chapter is to answer these questions and outline the overall design principles yielded by the answers.

6.1 Why Focus on Middle School?

In designing this curriculum, the first question that needed to be answered was: who is this curriculum for? What kind of student?

Due to the limited - but growing! - number of resources currently available to teach children about artificial intelligence, it was clear that the curriculum needed to be accessible to students with any kind of AI background, including no background.

The next question to follow was: what ages should the curriculum target? The middle school age range (approximately ages 10-14 years old) was decided on for two main reasons.

First, middle school is a good time to teach students about AI as it is the age when most children gain new technological independence. For example, the average age in which a child first receives a cell phone is 10.3 years old, or in the fifth grade. The average age in which a child first opens a social media account is 12.6 years old, or in the seventh grade [97]. Thus, introducing a curriculum focusing on the implications of AI-mediated technologies at this age has the potential for making a great impact in the way that these students consume technology for the rest of their lives.

Second, middle school is also a great time to introduce ethics because students in the middle school age range have higher levels of moral reasoning ability than their younger counterparts. As previously mentioned, according to Kohlberg's Theory of Moral Development, children in the middle school age range are capable of reasoning about conformity, authority, social order, and reciprocity [69]. We might expect younger students to have some difficulty with higher levels of moral reasoning, but we believe middle school students are a ripe age to synthesize both technical and ethical concepts.

6.2 Supporting Educators

While the curriculum is, of course, designed to teach children about the topic of AI and ethics, the materials themselves needed to be designed for educators to use. Educators, however, are a diverse set of individuals, from public school teachers to volunteers at after school programs, district initiative coordinators, and more. Educators can also have a variety of backgrounds - some may hold degrees directly related to ethics or technology; most will not.

In order to find out which educators (if any) were most interested in a curriculum on AI and ethics, I conducted a series of interviews with educators of various backgrounds. I interviewed one elementary school teacher from rural South Carolina, one middle school teacher from New York City, and one high school teacher from suburban Washington. Additionally, I interviewed coordinators for a public after school program in Somerville, MA, and a STEM Director for a school district in Pittsburgh, PA. Two of these educators held higher degrees related to computer science.

These interviews made it clear that there was a desire from a diverse group of educators to teach on the topic of AI. A common theme that recurred during these interviews was the need to prepare children for a future workforce where AI knowledge is seen as not only an advantage, but necessary. As one educator said, "This [teaching AI] is how we make our kids 'future ready.'" And it wasn't just educators in the traditional classroom setting who were interested in teaching their students about AI. As one after school director commented, "We're trying to change the way after school programs are viewed - it's more than daycare."

However, teachers also wanted more than "just" technical AI curricula - they also wanted ethics content. "I want my lessons to integrate a sophisticated understanding of identity," one teacher said, "and inspire them to not only be computer scientists, but also activists." As another educator said, "I think it's really important for kids to see that AI really affects all aspects of their lives."

Although initiatives exist to standardize AI education in the United States, these interviews made it apparent that teachers in the classroom - whether that be in the traditional, public school setting or in after school or workshop-like venues - drive change and progress in the space. These initiatives really are bottom up, grassroots movements from the teachers, not top down from the state or district level. In order for it to be the most effective, activities in the curriculum need to be easily accessible and understandable to educators.

6.3 Design Principles

Through interviews with K-12 educators (as well as the experts mentioned in previous chapters), the following design principles emerged and guided the development of the curriculum.

6.3.1 Accessibility

The largest concern, both from educators in the classroom setting as well as educators in the after school or summer camp setting, was budgeting and providing resources for an AI-centered curriculum. Many educators expressed fear over educational robotics kits being too expensive. All of the educators, however, mentioned having access to Chromebooks for students.

A secondary concern educators expressed was concern over teacher training, specifically, needing to learn new software or technical concepts. Educators were nervous about both the investment in time this might take on their behalf as well as the ever present concern that technology might malfunction and a lesson would become impossible to complete.

To address these concerns, I decided that the curriculum would be designed to be as "unplugged as possible." Unplugged activities are any activities that do not require technology (hence they are "unplugged").

Unplugged activities ensure that the curriculum is low-cost and is accessible to students, schools, and programs of all economic backgrounds. Additionally, unplugged activities are known to benefit students in many ways as the activities are often highly kinesthetic or constructivist, which offer a sense of play [30]. Many of the activities throughout the curriculum follow a constructivist approach where the teacher acts as a facilitator instead of lecturer and students often leverage their own schema on topics like YouTube.

In order to truly democratize AI, it is essential that as many barriers as possible be removed that would prevent low-income and underrepresented students (students who are at the highest risk of being negatively affected by AI systems) from understanding how AI works, the impact AI has on society, and the impact AI will have on society. For this reason, not only is the curriculum largely unplugged (requiring only pencil and paper; one lesson requires access to the Chrome web browser), it is also entirely open source and available at: https://bit.ly/mit-ai-ethics.

6.3.2 "I Do, We Do, You Do" & Leveraging What Kids Know

In order for the curriculum to be effective, it must be both accessible and engaging to students. Following the advice of interviewed teachers, many of the lessons follow the gradual release of responsibility model, also known as "I do, we do, you do." In this model, the instructor introduces the lesson topic, gives direct instruction by working through multiple examples, gives guided instruction as students work through examples as a class or in small groups, and then provides students the opportunity to complete the final portion of the lesson independently or in pairs.

This model, originally developed to teach reading comprehension strategies, has since been adopted across all content areas and is likely to be familiar to most educators. The model builds on known effective teaching strategies, including instructional scaffolding which allows instructors to differentiate scaffolding between students in order to keep them in the zone of proximal development, or the space where students cannot do an activity entirely by themselves but are capable of learning with assistance. Additionally, the gradual release of responsibility model allows hands-on learning to occur and builds student practice into the lesson [54]. Due to educators' familiarity with this model, it's similarity to the architecture of a mini lesson, another common teaching strategy, this model was adopted in many of the activities in the curriculum, or in the arc of the curriculum (for example, two activities included direct and guided instruction and a third activity was comprised of an independent task).

The gradual release of responsibility model emphasizes that each lesson begins with the teacher stating a connection to what students already know or have learned in a previous lesson. Connections ensure that a lesson is more meaningful to students and that they are more likely to learn during the lesson [54]. This sentiment was also echoed in interviews with teachers and parents, who urged that lessons relate to students' own experiences with technology. For this reason, a few lessons focus on writing algorithms to make a "peanut butter and jelly sandwich," a common introductory computer science activity, and design activities focus on YouTube, which according to Pew Center Research, is the most popular social media platform for young teens [23, 26].

6.3.3 Integrated Ethics Education

Finally, this curriculum sought an integrated approach to teaching ethics content, that is, that ethics content not be taught only after all technical content has been taught. As mentioned in Chapter 4, integrated approaches are gaining traction at the collegiate level, but work in this direction is still needed in the K-12 space [58, 39, 47, 60, 92].

Separating ethics modules from technical models is a problematic approach because research suggests that isolating ethics content often leads students to perceive ethics as unrelated to their technical studies or as a "side project." For this reason, each lesson in the curriculum focuses on some aspect of ethics and design [95, 46].

"Whoever codes the system embeds her views. Limited views create limited systems. Let's code with a more expansive gaze."

Joy Buolamwini

7 Pilot Workshop

In October 2018, I ran a pilot with over 125 children in grades 5th-8th grades at David E. Williams Middle School outside of Pittsburgh, PA. This study was approved by MIT's Committee on the Use of Humans as Experimental Subjects (COUHES), MIT's internal review board which reviews and approves research dealing with human subjects. David E. Williams Middle School is unique in that all of its students are beginning to learn about and design their own AI systems through courses such as *Introduction to Pattern-Finding through Gaming* and *Recognizing Computer Patterns Virtually and Through Algorithms* [27]. Artificial intelligence education has been integrated into all of their elective courses, and will soon make its way into their core curriculum as well.

At the time of the pilot, students had not yet begun any of these AI classes, although they recognized the importance of AI as it related to their future careers and knew they would be learning more about the technical underpinnings of AI later in the semester.

7.1 Protocol

Students participated in three lessons during their normal library elective period (which occurred every other day) which lasted approximately 45 minutes. Classes were organized by grade, with class sizes ranging between approximately 15-30 students each, with an average of 19 students per classroom. Students were given a pre-assessment and post-assessment at the beginning and end of each lesson, respectively, which asked questions about their learning for that class period, as well as surveyed students about their preconceptions of AI and the technology they used.

7.2 Lessons

Three activities were introduced during the pilot, one per class period, and are described below.

7.2.1 Lesson 1: Introduction to Supervised Machine Learning and Algorithmic Bias

The goal of this lesson was to introduce students to artificial intelligence, give them an idea of how AI technically works, and introduce them to the notion of algorithmic bias and the role of a training dataset in a supervised machine learning system.

The first lesson began with an introduction to artificial intelligence concepts and vocabulary. Students were asked to name examples of AI they were familiar with and were given a definition of supervised machine learning, which is a dataset that feeds into a learning algorithm that produces a prediction. Students were then introduced to the concept of classification as a form of pattern recognition and facial recognition as an example of classification.

Then, using Google's Teachable Machine webtool and printed images of cats and dogs, students were asked to train a cat-dog classifier. Unknown to them, the students were given a biased dataset. The dataset contains more cats than dogs and the collection of cat images were more diverse than the collection of dog images. Students were then given additional printed images of cats and dogs to test their classifiers and record their findings.

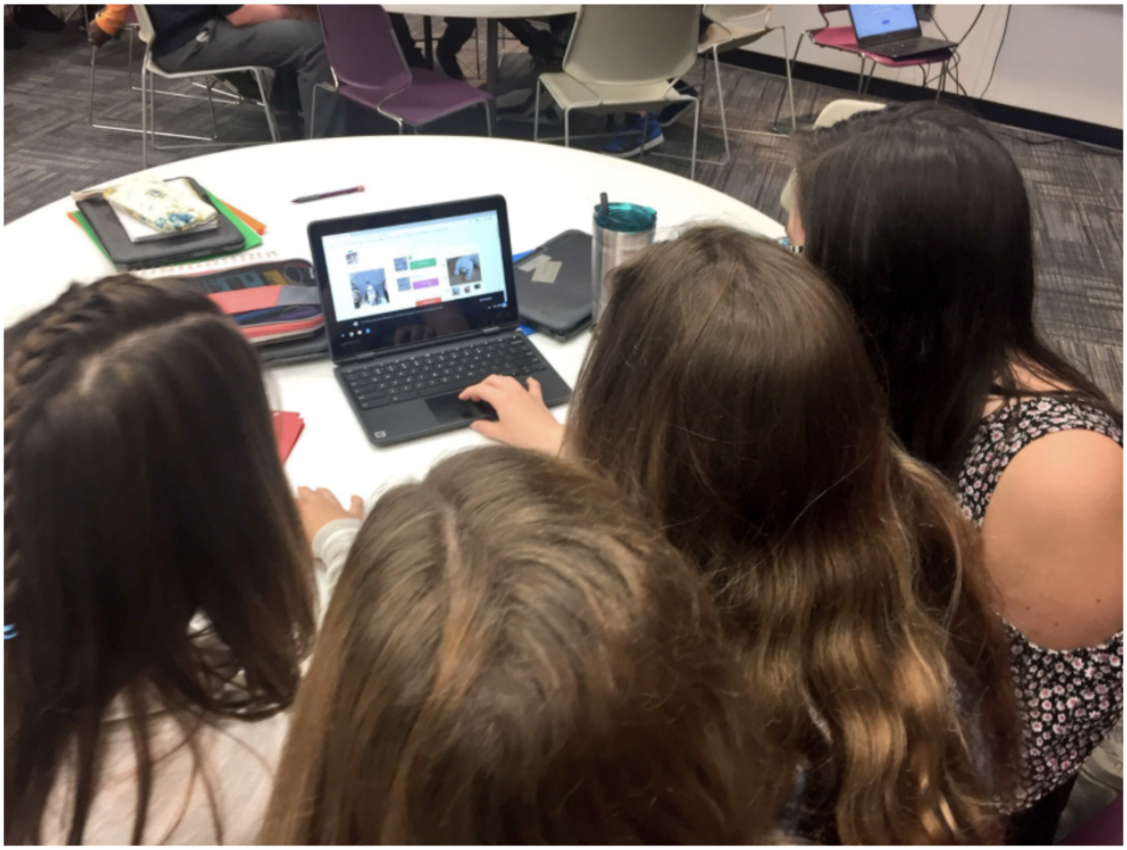


Figure 7-1: Students use Google's Teachable Machines tool to build their own classifier.

Once students found that the classifier works better on cats than dogs, they had the opportunity to retrain their classifiers with a new dataset. The lesson concluded with students being shown Joy Buolamwini's YouTube video titled "Gender Shades" which reveals that commercial facial recognition systems work better on lighter skinned male faces than darker skinned female faces [35]. Students were prompted to discuss how Buolamwini's findings connected to the activity they just completed.

Figure 7-2: Students watch as scholar Joy Buolamwini explains her research showing that commercial facial recognition systems work better on lighter skinned male faces than darker skinned female faces.

7.2.2 Lesson 2: Algorithms as Opinions and the Ethical Matrix

The goal of this lesson was to introduce students to the idea that algorithms are not neutral; rather, they are more like opinions. Additionally, students were introduced to the concept of stakeholders and learned about the ethical matrix as a way to reason about stakeholders and their values in a technical system.

The second lesson opened with a discussion about what students learned in the previous lesson. Students were asked to recall the meanings of the following vocabulary: dataset, prediction, and classification. They were then introduced to the concept of an algorithm in comparison to a recipe, as something that takes in an input, has specific steps to modify that input, and produces an output. Students then dug into the concept of a “learning algorithm” and discussed that learning algorithms try to find patterns in a given dataset. For example, in the previous lesson the algorithm was looking for similarities and differences between images. However, there are other kinds of patterns an algorithm could find. For example, a robot might try to learn what makes the “best” peanut butter and jelly sandwich by looking for patterns in feedback between different iterations of the sandwich.

Building on a popular activity in both elementary computer science classrooms as well as elementary language arts classrooms, students were then asked to write out specific instructions for their own “peanut butter and jelly sandwich algorithm” [23].[1] Students needed to specify which “input” were necessary (ingredients as well as kitchen utensils) as well as write out specific instructions to produce the desired output.

After writing their specific sets of instructions, students were then asked to describe the result of their “best peanut butter and jelly sandwich” algorithm. Students were introduced to the term “optimization” and then asked to reflect on what goal their PB&J algorithm was optimizing. Most commonly, students optimized for the tastiest sandwich by writing instructions to cut off crust or substituting grape jelly for another kind of jam. This reflection led into a discussion about what other goals their PBJ algorithm could have optimized for: perhaps the algorithm could have optimized for the sandwich that took the least amount of time to make? Or the healthiest sandwich? This line of questioning served to show students that their algorithms reflected their opinions about what makes the “best” sandwich.

Students were then posed with the following question: how do we decide which goal our PB&J sandwich should have? Students were then introduced to the concept of a stakeholder and asked to discuss who stakeholders might be in their PB&J sandwich

[1] This activity was also inspired by Cathy O’Neil’s video “The Truth About Algorithms” where she uses a similar example - that of preparing a meal for her family - to show that algorithms essentially represent the conscious or unconscious opinions of their makers. The video can be found here: https://vimeo.com/295525907.

algorithm: parents, siblings, doctors, dentists, grocers, and of course, themselves!

Afterwards, students constructed ethical matrices around the technology of self-driving cars. They identified several values stakeholders could have in an autonomous vehicle, such as safety, efficiency, accessibility, and others. They also brainstormed as a class and then in small groups about stakeholders related to self-driving cars, such as passengers, pedestrians, car companies, and the government.

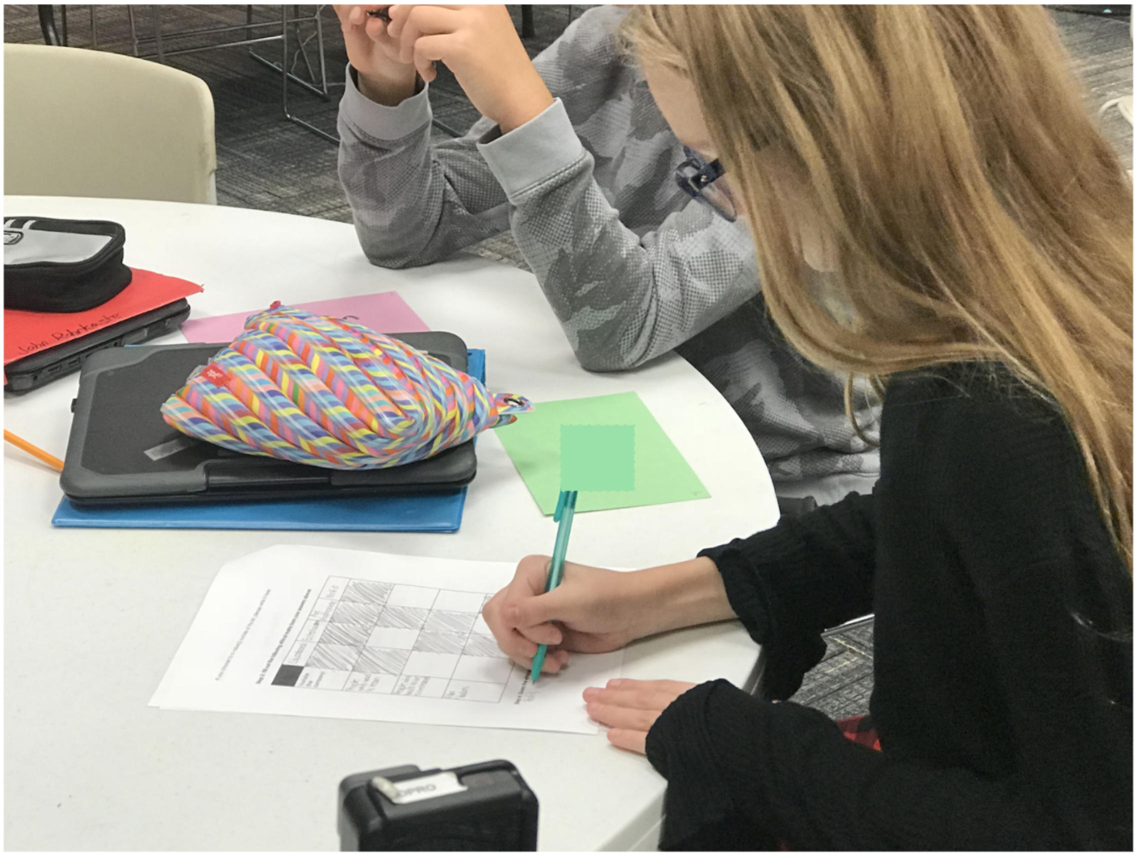


Figure 7-3: Students complete an ethical matrix around self-driving cars in pairs.

7.2.3 Lesson 3: YouTube Redesign

The goal of the third and final lesson was to have students apply what they had learned. Students were tasked with "redesigning YouTube" by identifying the stakeholders invested in YouTube's recommendation algorithm (or other algorithms, such as the comment filtering algorithm or the algorithm which supplies search results) and the values those stakeholders may hold in the system. Students were then asked to build an ethical matrix around these stakeholders and values and use the ethical matrix as a conversation starter to determine a new goal - a new “opinion” - of the recommendation system.

Then students were prompted to consider what kinds of data their algorithm would need to learn to achieve this goal and what kinds of data they would need to collect to teach their algorithm.

At this point in the activity, students paused and were shown a demo of Gobo, a platform which redesigned Twitter, and engaged in a conversation about the purpose of various features on a platform [15]. For example, students recognized that a “like” button not only served as a social signal to their network but also acted as a data point from which a recommendation algorithm could learn.

Finally, students were given craft supplies and asked to paper prototype in groups of 3-5 students a YouTube interface which better reflected the values of their reimagined platform.

7.3 Results and Discussion

Below we discuss findings from this pilot study. All findings reported below are based on the number of students who participated in the activity in question. Pre-assessment surveys about students’ prior knowledge of AI and ethics were taken during the first class period before the first activity.

7.3.1 What Students Already Know about Ethics and AI

Course Expectations

At the beginning of the pilot, students were asked both what they expected to learn during the week as well as some questions about their technology usage; 102 students answered this question. These questions can be found in Appendix C. It is worth noting that the majority of students had not yet begun the other AI courses, and therefore the majority of students had no previous programming or AI experience. A handful of students (5-10 students) verbalized previous AI experience with tools like Cozmo or Lego Mindstorms. We found that students' expectations for what they would learn in a course entitled "AI and Ethics" varied widely. The most common answer to this question was simple: I don't know. When asked, 28.4% of students answered this question with some variant on "idk." Beyond these students, many students stated an expectation to learn technical concepts about artificial intelligence. In their responses, 22.5% of students' responses mentioned terms like engineering, programming, coding, or phrases like "how it works" or "how it is made." One student wrote, "I hope to learn the engineering and functioning of AI. Some questions I have are how do you program AI to make it think or process?" Another student wrote, "I expect to learn how to program AI. I want to learn how to create an AI."

Similarly, a common theme among students' expectations was that of learning about robots. The term "robots" was included in 15.6% of students' responses. The majority of these responses were similar to the responses above: almost all students who expected to learn about robots indicated they expected to learn how robots "work" or "are programmed" or "function."

Beyond these responses, a few students indicated that they were aware of larger questions of "ethics" around AI. One student wrote that they expected to learn, "How the AI world will effect [sic] the real world today." Another wrote, "[I expect to learn] how AI works, how it affects our day to day lives in society." One student specifically mentioned morality in the context of AI, writing, "I expected to learn what we could do with AI, and which of those things is morally right and why," while another identified a key question of who AI helps, writing, "I expect to learn about AI, what it does, and who/what it helps."

Technical Understanding of AI

In addition to these expectations, we were able to learn about which technologies children interact with regularly, and of these technologies, which of them they perceive to use AI. The most popular technologies children listed interacting with regularly included: Instagram, Snapchat, Google Search, YouTube, and Spotify. Students were less familiar with technologies like Facebook or Twitter. However, when asked to name applications that utilized artificial intelligence, 32.3% of students were unable to do so. Of the students who responded, 46.3% of students listed only voice agent devices such as Siri, Cortana, Google Home, and Amazon Alexa. Seven students listed Google Search as an application of AI, and only four students listed YouTube.

When asked how AI works, the majority of students, 53.9% of them, reported that they didn't know or were unsure. Those who gave responses primarily described the interaction design of voice agent technologies. One student wrote, "Siri, Hey Google, and Alexa are all devices you can talk and interact with. For example if I were to say 'Alexa, put on Anne-Marie,' she would start to play music by Ann-Marie." Another student wrote, "if you say a name the AI will turn on, if you ask a question it will turn on and answer." In fact, many students described question-asking as inherent to AI, writing, "it helps people answer questions and it helps suggest things to look up" or "they all help you with questions you have", and "You can ask it a question and it will search [for] an answer for you." A few students acknowledged that AI is programmed and created by humans, writing, "It is programmed by humans to make decisions about how to do its job based on its situation" or "Someone programs the software and it makes life easier". Additionally, two students referenced internet connectedness, stating "this AI builds its own code to provide an answer from the internet" or "you would start a conversation with it then it would search the internet with a [sic] answer." One student also showed some understanding of sensors and rules-based AI, writing, "AI is a robot using sensors in feel, sight, and sound to produce answers and "if statements."

Ethical Understanding of AI

The question "who is affected by AI?" was answered by 111 students. The most common answer among students was an all-encompassing response like "humans" or "everybody" or "people," with 56.7% of students responding this way. When asked, 23.4% of students were unable to name any group or party affected by AI. A small percentage of students, 11.7% of them, mentioned more specific stakeholders such as technology companies, workers, scientists, or their family. The remaining $\tilde{9}$% of students gave answers that did not make sense in the context of the question, such as non-human answers like "machinery" or questions like "what is it?".

Finally, at the beginning of the pilot, students were asked how AI can be "good" or "bad." When explaining how AI can be good, the primary keywords students used were "help" and "easier." Many student responses began with "It can help. . . " and wrote about how AI could assist with learning, such as "it could help you learn" or "it can teach and help people study" or "it can help with math." While the majority of responses described how AI could be useful in terms of education and learning, a minority of other divergent answers appeared, such as, "it can cure cancer" or "it can make jobs easier for humans" or "it can help people with a disability still live." One student wrote, "AI can be good because society use [sic] it to clean up toxic war zones or dangerous places humans cannot go."

When it came to discussing the potentially harmful effects of AI on society, student responses were more varied than those focusing on the benefits of AI. Student responses tended to focus on the following four topics: the idea of AI taking over the world, privacy concerns, the impact of automation, and general concerns about technology not functioning as intended. The idea that AI could become too powerful was the most popular topic of concern. Students wrote responses like "yes it can take over humanity" or "robots can take over the world" or "it can be too powerful." Some students shared privacy concerns, writing responses such as, "It can spy on us" or "it can be use [sic] to steal money and see where your address is." Other students showed concerns about the role of AI in the economy and the impact of automation. A few students wrote, "a lot of people will lose your [sic] job" or "it could be bad because it could overpower jobs and make people earn less money." A few students just had general concerns related to the idea of technology malfunctioning, writing, "AI can be bad because it can malfunction and mess up" or "I believe that AI can bre [sic] bad because there are always kinks in the technology therefore meaning that it can malfunction" or "yes, scripts can fail and code can glitch." The majority of students identified ways AI could both benefit and harm society.

7.3.2 What Students Learned

Technical Understanding of AI

After interacting with AI and building their own models during the first activity, students were again asked to identify examples of AI. At this point in time, 4.3% of students were unable to do so (down from 32.3%). Additionally, student responses were more diverse than the answers they gave before the workshop began. Voice agent technologies like Siri or Google Home were still the most frequently occurring examples provided, but only 11.5% of students listed only these technologies as their examples of AI (down from 46.3%). In addition to voice agents, 28.9% of students listed facial recognition technology as an example of AI. This is unsurprising given that facial recognition is discussed heavily during the first activity. Beyond facial recognition technology, 13 students listed Snapchat as an application of AI (again, likely because Snapchat filters were discussed during the first activity), 7 students listed Google Search as an application of AI, and 5 students listed YouTube.

When asked how AI works after the first activity, some students continued to describe the interaction design of voice agents. In 18.8% of responses, the keyword "question" was included, always in the context of asking a voice agent a question. Furthermore, 26.0% of students made a reference to AI using "data" or a "dataset" as past examples to learn from to make predictions. One student wrote, " [AI] uses a dataset to make a prediction." Another student wrote more specifically, "AI learns from a provided dataset and a learning algorithm are used [to] program AI to predict." Some students applied this knowledge to a technology they are familiar with when explaining how AI works, "With youtube [sic] recommendations come up after you watched similar videos or if you watched on person's video another might come up."

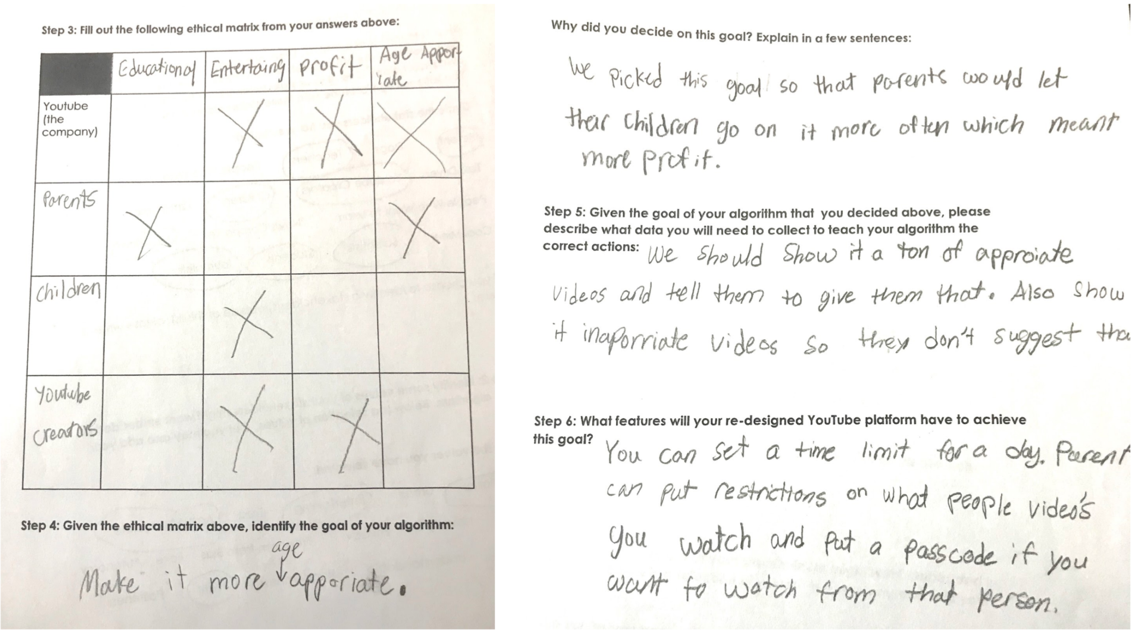
Ethical Understanding of AI

Students showed a capacity to engage with the ethical material at varying levels. When it came to the ability to identify who is affected by artificial intelligence, student responses were very similar to the pre-survey results. For example, 62.3% of students gave a general answer like “humans” or “everybody,” while 20.5% of students were unable to identify any stakeholders. The number of students who were able to list a few specific stakeholders did increase, from 11.7% in the pre-survey to 17.0% in the post-survey. Of these specific stakeholders, “workers” and tech companies (often written as “Google” or as “people who make it [AI]”) were the most common. A few students did write responses such as “darker colored women, pedestrians, drivers, passengers in a car,” indicating they were reflecting on the activities completed beforehand in writing their answers.

When asked how AI could be bad, students shared sentiments like “I learned that some AI, like facial recognition, has trouble identifying darker people and females," showing that they understood the finding in Buolamwini and Gebru’s work as unfair. Many more students, however, responded to this question with concerns about the same narratives discussed at the beginning of the workshop. One student cited concern over the future of work, “Yes, it can take jobs from humans," as well as concern about the robot apocalypse, writing, “Also, there’s terminator."

Example Projects

Students’ paper prototype projects also gave insight to their ability to apply what they learned throughout the workshop. For example, the project shown in Figure 7-4 identified parents, children, and YouTube creators as stakeholders in addition to YouTube. Even though “entertaining” was the value that the most stakeholders had in common, this group decided to optimize their YouTube Redesign for “age appropriateness.” They justified this goal by writing, “We picked this goal so that parents would let their children go on it [YouTube] more often which meant more profit.” When asked how they will “teach” their algorithm to optimize for this goal, students wrote, “We should show it [the algorithm] a ton of approiate [sic] videos and tell them [the algorithm] to give them them [users] that. Also show it inapporriate [sic] videos so they don’t suggest tha [sic].”



Step 3: Fill out the following ethical matrix from your answers above:

	Educational	Entertaing	Profit	Age Apporrate
Youtube (the company)		X	X	X
Parents	X			X
Children		X		
Youtube creators		X	X	

Step 4: Given the ethical matrix above, identify the goal of your algorithm: Make it more age apporiate.

Why did you decide on this goal? Explain in a few sentences: We picked this goal so that parents would let their children go on it more often which meant more profit.

Step 5: Given the goal of your algorithm that you decided above, please describe what data you will need to collect to teach your algorithm the correct actions: We should show it a ton of approiate videos and tell them to give them that. Also show it inaporriate videos so they don't suggest tha

Step 6: What features will your re-designed YouTube platform have to achieve this goal? You can set a time limit for a day. Parent can put restrictions on what people video's you watch and put a passcode if you want to watch from that person.

Figure 7-4: An example of students’ ethical matrix and justification for their redesign.

This group’s prototype added one new feature to the already existing YouTube interface: a flag denoting whether videos are age appropriate for the logged in user or not, seen in Figure 7-5.

A different group also implemented a design for “age appropriate” YouTube, seen in Figure 7-6. Students displayed content in different visual sections based on age and length of videos, with visible control options.

Another group optimized their YouTube redesign around “politeness,” seen in Figure 7-7. Instead of focusing on the recommendation algorithm, this group chose to focus on the algorithm which displays comments beneath videos. These students decided to visually flag potentially offensive comments while also having visible “restriction options” for users to change what is displayed with the specific ability to change the flags or to remove comments altogether.

One group decided to optimize their version of YouTube to be “less addictive.” Their project, shown in Figure 7-8, shows the recommendation algorithm recommend-

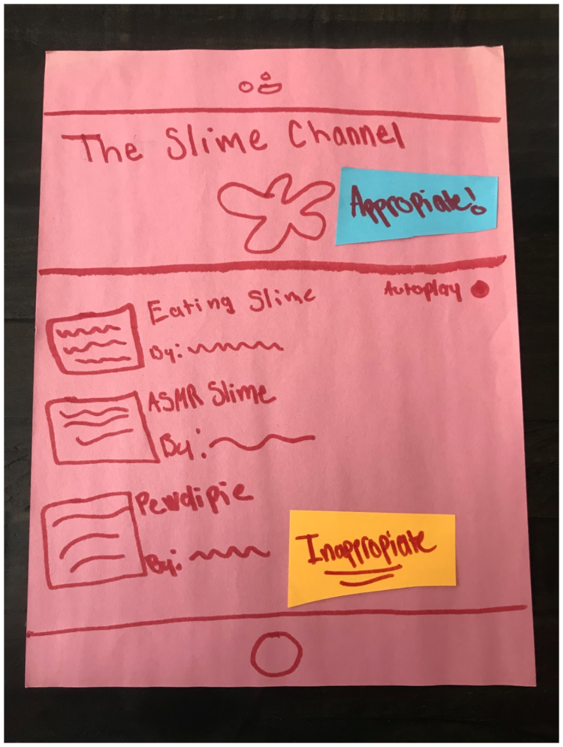


Figure 7-5: A paper prototype of "age appropriate" YouTube.

ing progressively "cringier and cringier" content. A user might start off by watching videos of dogs, but after a certain amount of time the algorithm begins recommending less pleasant videos like "nail on a chalkboard" or "picking your nose." The students also decided that the display would change over time from YouTube's standard white and red interface to a less visually pleasing interface dominated by brown tones.

At the end of the pilot, one student wrote, "It [designing AI for good] means to enhance products for the people, making activities easier or less work. Also, to make stakeholders happy."

Course Reflections

At the end of the course, students recorded which activity they learned the most from, which activity they found to be the most fun, and any feedback they had on the activities as a whole. As they reflected on the course, 54.2% of students said they learned the most from activity #2, which introduced the ethical matrix, followed by 26.5% saying they learned the most from activity #3, the YouTube Redesign activity.

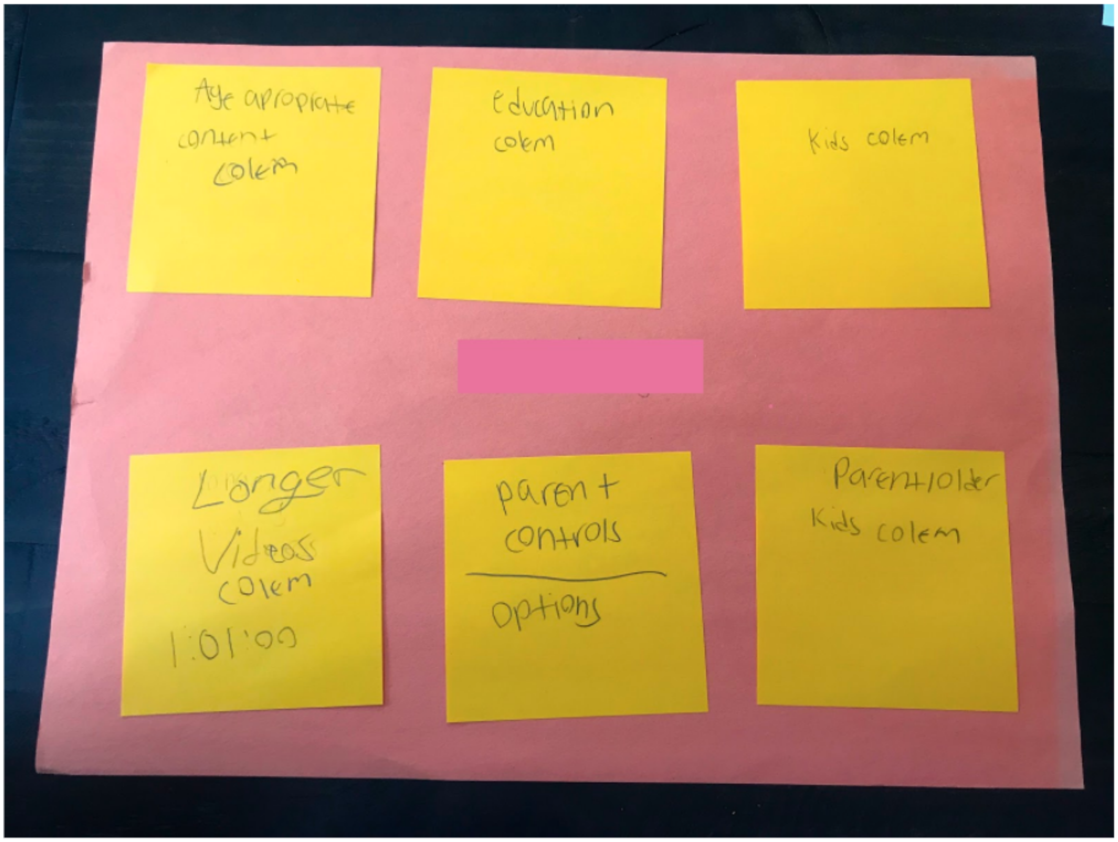


Figure 7-6: Another paper prototype of "age appropriate" YouTube.

Only 19.2% of students said they learned the most from the first activity which used Teachable Machines. Overwhelmingly, 66.6% of students said YouTube Redesign was the most fun activity, 19% of students said the ethical matrix activity was the most fun, followed by 14.2% of students who reported that the Teachable Machines activity was their favorite.

Feedback from students was overall very positive. The most common keyword was "fun," used by 37.5% of students who gave feedback. Two students indicated they did not enjoy the course but did not indicate a specific reason why. One student wrote, "It was quite boring," while another wrote "I don't like it. It sucks." A few students gave more specific feedback. One student wrote, "I enjoyed the examples from day 2, they were interesting and easier to understand," while other students indicated they wanted more time to engage with the material. One student wrote, "It was fun but a waste of time to only have three classes," and another suggested, "I think it should

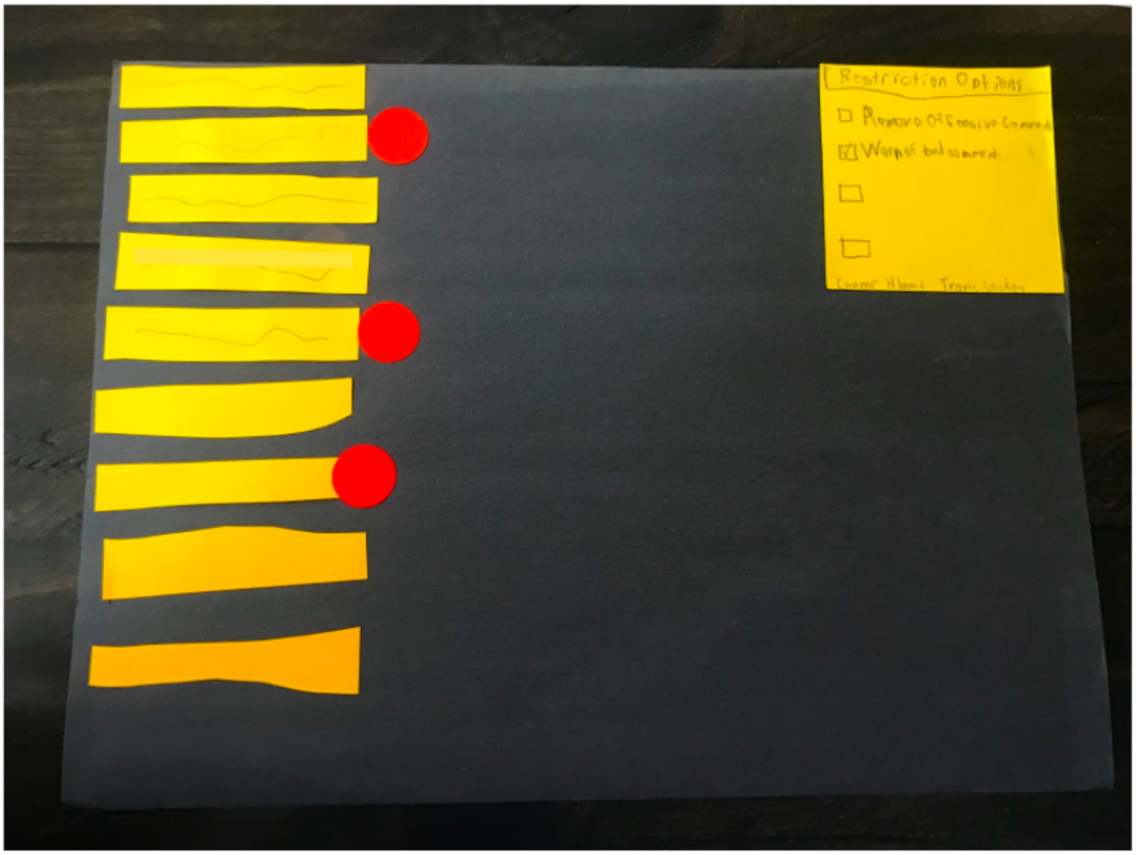


Figure 7-7: A paper prototype of "polite" YouTube.

happen at the end of the day where there is more time."

7.4 Curriculum Design Recommendations

From the pilot, I was able to learn which practices were successful and where areas for improvement existed for the next design iteration, all of which I discuss below.

7.4.1 Best Practices

As the student above remarked, one best practice that emerged from the pilot is the usefulness of a variety of examples in helping students formulate patterns and learn. This is unsurprising given the "I do, we do, you do" model mentioned in the previous chapter, which hinges on students having multiple opportunities, and therefore

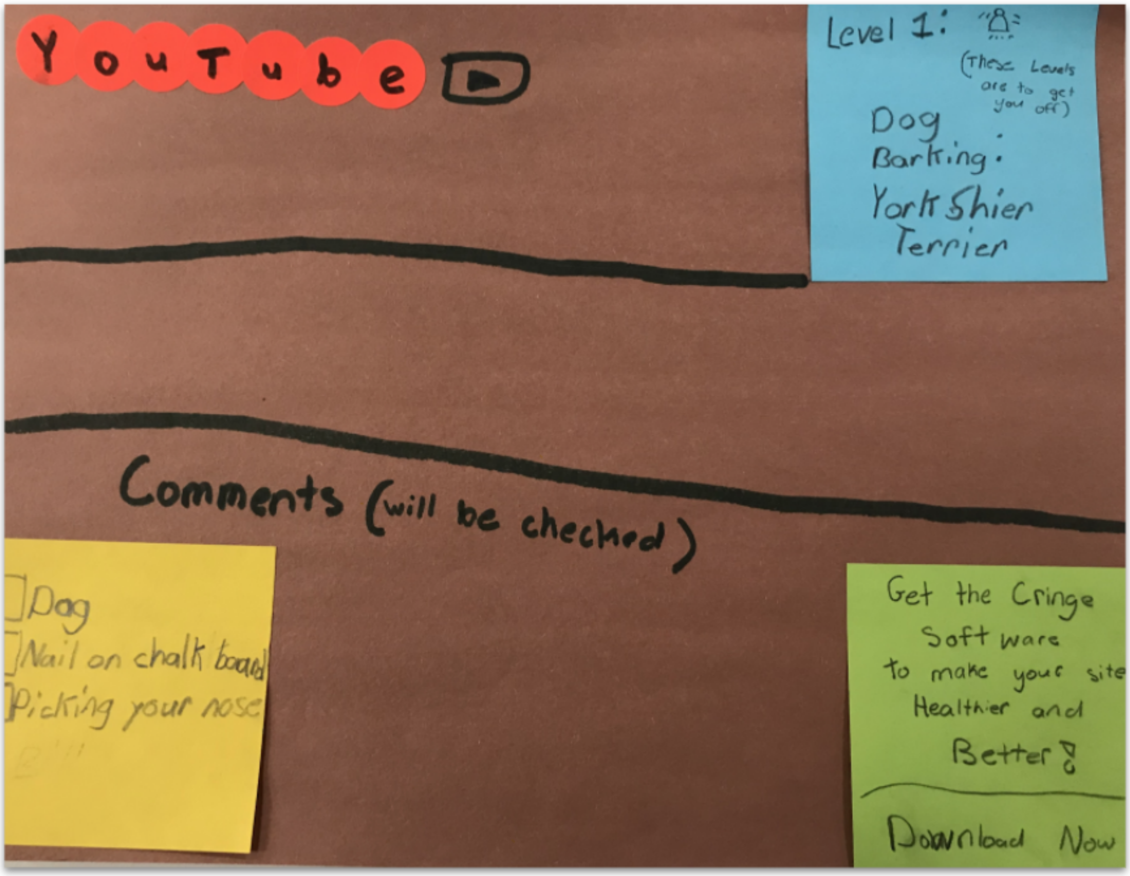

YouTube
Level 1:
(These Levels are to get you off)
Dog Barking:
York Shier Terrier
Comments (will be checked)
□Dog
□Nail on chalk board
□Picking your nose
Get the Cringe Soft ware to make your site Healthier and Better!
Download Now

Figure 7-8: A paper prototype of "less addictive" YouTube.

examples, of a technique or content. When discussing applications of AI, students did better when a variety of diverse examples were presented - both fictional and nonfictional AI, as well as embodied and non-embodied examples. Students also used the ethical matrix as more of a conversation starter (as opposed to majority takes-all voting scheme) when multiple examples were presented in class and discussed. Instead of solely seeing which value was the most popular among stakeholders, students discussed if some values were categorically more important than others in their projects because several examples of ethical matrices were discussed in different ways as a class together.

Similarly, students showed the most mastery of content when it was related to their own lived experience. Again, this is unsurprising given the "I do, we do, you do" model, which suggests that every lesson should begin with a connection. However, students seemed to engage more with the material when discussion centered on technology which situated them as experts. For example, when discussing facial recognition technology, students did not connect as much with examples of surveillance technology, say, used in airports. However, students discussed the potential harms of facial recognition technology and algorithmic bias more enthusiastically when it was connected to Face ID on their iPhones or Snapchat filters - technologies they use daily. Additionally, it appeared that students were able to identify more stakeholders during the YouTube redesign activity than during the activity about self-driving cars. In the future, connecting content to technology students already use seems to be an essential way to engage them in learning.

Finally, as evidenced by students rating the second activity as the most informative, it seems that engaging students in ethical design by invoking specific protocols (such as the ethical matrix) is another best practice. Although students enjoyed discussing the Heinz dilemma, it would seem that students did not find the abstract discussion as informative as the tool. In the future, activities should have students follow specific actions or prompts to engage in ethical design and thinking.

7.4.2 Content Recommendations

A few areas of improvement were also identified based on this pilot study's results, and recommended changes to the activities were identified. These changes come in two forms: changes to the curriculum's content as well as changes to the curriculum's structure.

Regarding content, it is clear that more time and emphasis should be spent on helping students understand and identify examples of artificial intelligence in the world around them. At the beginning of the workshop, the majority of the students conflated AI with robots or voice agents. They also did not seem to have a grasp on the difference between terms like programming, hardware, electronics, algorithm, which could help them better understand how AI works. This content should be included in future designs of the curriculum.

Similarly, it was evident that the majority of students who had preconceived notions about the "ethics of AI" were concerned with a few dominant media narratives of AI. Specifically, they were concerned about narratives around AI "taking over the world" as well as the impact of automation, and general concerns about privacy. Future iterations of the curriculum will need to address the reality of these concerns.

7.4.3 Structure Recommendations

Regarding structure, many students voiced feeling rushed for time. In the future, longer or more class sessions are recommended, especially for the YouTube Redesign activity. Additionally, it is worthwhile to rethink group structures or the process for choosing stakeholders in the YouTube Redesign activity. Many larger groups (groups with four or five students) struggled to come to a consensus on which stakeholders to include. It might be more time efficient to have students choose from a bank of stakeholders or to have them work in partners instead of groups for the sake of coming to a timely consensus (especially when there are few instructors to help facilitate individual group discussions).

Overall, it is evident that students have a potential to engage and understand with the core concepts. Future iterations of these activities which provide students with more foundational knowledge, such as differentiating AI from robots, as well as more time to work on projects, will likely assist in their learning.

"That's the kind of design thinking I hope and wish for: Where 'what's wrong' drives our pursuit of 'what if?'"

Sherri Spelic

8

Standards and Curriculum

8.1 Overview

Based on the pilot study in Pittsburgh, the following set of standards were devised to guide what children should know about artificial intelligence - both how it works and its impact on society. The curriculum and activities were also revised to reflect these standards and the design recommendations learned in Pittsburgh. Below, the standards are presented, and the larger curriculum is described.

8.2 Another Note on Terminology

Before delving into the suggested standards, I would like to offer a note about the terminology contained within them, which differs slightly from the rest of this thesis. In several places the phrase "socio-technical system" appears where the phrases "artificial intelligence" or "algorithm" might be expected.

The phrase "socio-technical system" originated in the field of system design and is used to highlight the fact that technologies do not operate in vacuums; rather, they are embedded in social systems and their success relies on how those technologies are embedded in these social systems. For example, platforms like Twitter display trending topics to their users. Twitter decides which topics are trending not based on the frequency of any given phrase, but based on how quickly a phrase is gaining popularity. That is to say a phrase such as "snowday" could have a lower frequency than another phrase like "tbt," (an abbreviation for the phrase "throwback Thursday" which is often used weekly millions of Twitter users). However, if in a particular time span "snowday" is mentioned at a higher rate than "tbt," Twitter will show "snowday" as trending instead of "tbt." It's easy to think of this algorithm as a simple mathematical equation - calculate the rate of change in keywords and the output is a trending topics feature. However, the system is not just technical - it is socio-technical. As soon as a topic trends, a snowball effect occurs. Since users see the topic as trending, they tweet with it more, and thus, it continues to trend, which can lead users to thinking a news event is more important to their country than it actually is, can lead to journalists reporting on this topic and making it a bigger issue, and so on. The trending topics algorithm is not simply a technical system, rather it is informed by and informs society in a large, tangled web of loops and feedback loops.

It is important to note that even technologies without direct or obvious societal impact (or "real world application") are also socio-technical systems. For example, a piece of software which is built solely to simulate theoretical mathematics is still a socio-technical system; it is still affected and affects society. Those who develop the software bring to the project a culture and practices from previous projects which determine how the codebase is structured, which functions are prioritized, who is able to work on which pieces. And the culture and practices they establish around this piece of software carry into the next piece - and eventually lead to norms around building software which can be beneficial or harmful to users or other developers.

The phrase is used intentionally below in the standards to highlight, particularly to teachers, that AI systems are actually socio-technical systems. Furthermore, the phrase is used to highlight one of the benefits of teaching ethics from a design perspective: many of these standards apply to artificial intelligence but they may also be easily transferred to other areas of computer science or engineering education. Many of the standards could be applied to cybersecurity, web design, or even civil engineering curricula - all of which focus on socio-technical systems. In the standards below, “artificial intelligence” and “algorithm” are used when learning objectives are specific to those topics and “socio-technical system” is used in standards that could be easily transferred to other areas of study.

8.3 Suggested Standards

The following are the standards I developed after the initial pilot for the second version of this curriculum (described in the next section). These standards were developed with middle school-aged students in mind.

1. Understand the basic mechanics of artificial intelligence systems.
   (a) Recognize algorithms in the world and be able to give examples of computer algorithms and algorithms in everyday contexts (for example, baking a cake).
   (b) Know three parts of an algorithm: input, steps to change input, output.
   (c) Know that artificial intelligence is a specific type of algorithm and has three specific parts: dataset, learning algorithm, and prediction.
   (d) Recognize AI systems in everyday life and be able to reason about the prediction an AI system makes and the potential datasets the AI system uses.
2. Understand that all technical systems are socio-technical systems. Understand that socio-technical systems are not neutral sources of information and serve political agendas.

    (a)

3. Recognize there are many stakeholders in a given socio-technical system and that the system can affect these stakeholders differently.

    (a) Identify relevant stakeholders in an socio-technical system.
    (b) Justify why an individual stakeholder is concerned
    (c) about the outcome of a socio-technical system.
    (d) Identify values an individual stakeholder has in an socio-technical system, e.g. explain what goals the system should hold in order to meet the needs of a user.
    (e) Construct an ethical matrix around a socio-technical system.

4. Apply both technical understanding of AI and knowledge of stakeholders in order to determine a just goal for a socio-technical system.

    (a) Analyze an ethical matrix and leverage analysis to consider new goals for a socio-technical system.
    (b) Identify dataset(s) needed to train an AI system to achieve said goal.
    (c) Design features that reflect the identified goal of the socio-technical system or reflect the stakeholder's values.

5. Consider the impact of technology on the world.

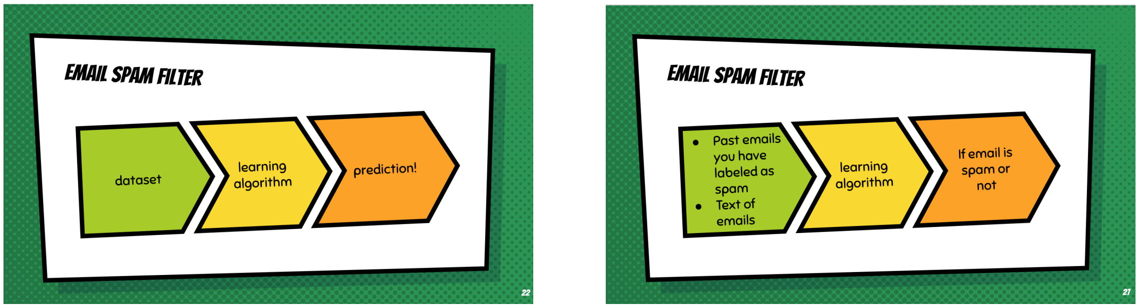
    (a) Reason about secondary and tertiary effects of a technology's existence and the circumstances the technology creates for various stakeholders.

8.4 Curriculum Activities

Below are the eight activities that were developed for the final version of the curriculum that were piloted during a weeklong summer workshop (described in Chapter 9). All materials and teacher guides for the activities can be found at www.bit.ly/mit-ai-ethics.

8.4.1 AI Bingo

In this activity, students are introduced to the definition of artificial intelligence. That is, they learn that a type of AI, supervised machine learning, is comprised of three parts: a dataset, a learning algorithm, and a prediction. The instructor then presents a few examples of an AI system according to these definitions. As a class, students brainstorm examples of AI according to the definition by identifying a technology possibly powered by artificial intelligence, the prediction that technology is trying to make, and a dataset it uses to make said prediction.



EMAIL SPAM FILTER

dataset → learning algorithm → prediction!

22

(a) Three parts of an AI system.

EMAIL SPAM FILTER

- Past emails you have labeled as spam
- Text of emails

→ learning algorithm → If email is spam or not

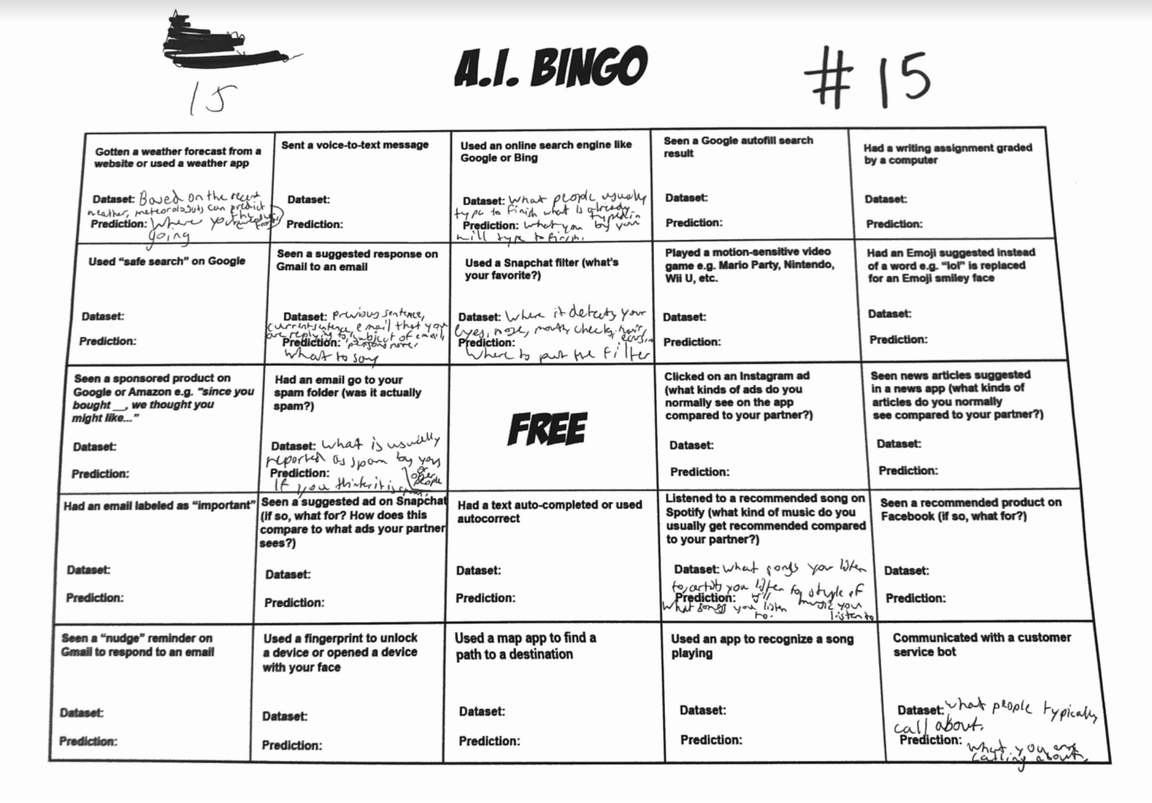
27

(b) Labeled Spam Filter

Figure 8-1: An example of an AI system given in class. For the purposes of this class, all AI systems are broken down into three parts: a dataset, learning algorithm, and prediction.

After this short class discussion, students are given bingo cards with various AI systems. Students are tasked to find a partner who has also used that AI system and together they work to identify what prediction the system is making and the dataset it uses to make that prediction.

This activity was included for two reasons. First, the activity was included in response to the pilot results which indicated students needed more practice with the definition of artificial intelligence and more practice identifying non-embodied and non-voice agent examples of AI. Second, the activity was included as an initial icebreaker exercise so that students in the workshop could meet their peers and have a low-pressure way of interacting with each other.



[redacted]

15

A.I. BINGO

#15

Gotten a weather forecast from a website or used a weather app Dataset: Based on the recent weather, meteorologists can predict the weather forecast Prediction: Where you're going	Sent a voice-to-text message Dataset: Prediction:	Used an online search engine like Google or Bing Dataset: What people usually type to finish what is already typed in Prediction: what you will type to finish.	Seen a Google autofill search result Dataset: Prediction:	Had a writing assignment graded by a computer Dataset: Prediction:
Used "safe search" on Google Dataset: Prediction:	Seen a suggested response on Gmail to an email Dataset: previous sentence, current sentence, email that you are replying to, subject of email, persons name, Prediction: What to say	Used a Snapchat filter (what's your favorite?) Dataset: Where it detects your eyes, nose, mouth, cheeks, hair, ears, [illegible] Prediction: Where to put the filter	Played a motion-sensitive video game e.g. Mario Party, Nintendo, Wii U, etc. Dataset: Prediction:	Had an Emoji suggested instead of a word e.g. "lol" is replaced for an Emoji smiley face Dataset: Prediction:
Seen a sponsored product on Google or Amazon e.g. *"since you bought __, we thought you might like..."* Dataset: Prediction:	Had an email go to your spam folder (was it actually spam?) Dataset: What is usually reported as spam by you or other people Prediction: If you think it is spam.	FREE	Clicked on an Instagram ad (what kinds of ads do you normally see on the app compared to your partner?) Dataset: Prediction:	Seen news articles suggested in a news app (what kinds of articles do you normally see compared to your partner?) Dataset: Prediction:
Had an email labeled as "important" Dataset: Prediction:	Seen a suggested ad on Snapchat (if so, what for? How does this compare to what ads your partner sees?) Dataset: Prediction:	Had a text auto-completed or used autocorrect Dataset: Prediction:	Listened to a recommended song on Spotify (what kind of music do you usually get recommended compared to your partner?) Dataset: what songs you listen to, artists you listen to, style of music you listen to Prediction: what songs you'll listen to.	Seen a recommended product on Facebook (if so, what for?) Dataset: Prediction:
Seen a "nudge" reminder on Gmail to respond to an email Dataset: Prediction:	Used a fingerprint to unlock a device or opened a device with your face Dataset: Prediction:	Used a map app to find a path to a destination Dataset: Prediction:	Used an app to recognize a song playing Dataset: Prediction:	Communicated with a customer service bot Dataset: what people typically call about. Prediction: what you are calling about.

Figure 8-2: An example of a student's AI bingo card.

8.4.2 Algorithms as Opinions

The next two activities build on Activity #2 from the pilot workshop. In this activity, students are presented with the definition of an algorithm as an input, steps to change the input, and an output. This definition of an algorithm is also related to the definition of AI presented earlier as input is mapped to a dataset, steps to change the input are mapped to a learning algorithm, and the output is mapped to a prediction. The instructor compares an algorithm to a recipe for baking a cake and maps the input to ingredients, steps to change the input to mixing ingredients and baking the mixture in the oven, and an output - a cake!

With this example in mind, students are then asked to write an algorithm to make the "best" peanut butter and jelly sandwich. They name the ingredients required for their sandwich, the specific steps needed to make their sandwich, and a description of the final output itself. Afterwards, students share their algorithms with the rest of the class. In doing so, they find that there are many definitions of "best." Most students state that they are trying to make the most delicious sandwich, but find that they try to achieve this goal in very different ways. For example, some students opt for grape jelly and others opt for strawberry jam. Some students cut off the crust of their PBJ and some skip the peanut butter entirely. The instructor then asks why “best” has to mean “most delicious.” Why can’t “best” mean the healthiest, quickest to make, or the cheapest? Students discuss these goals as a class.

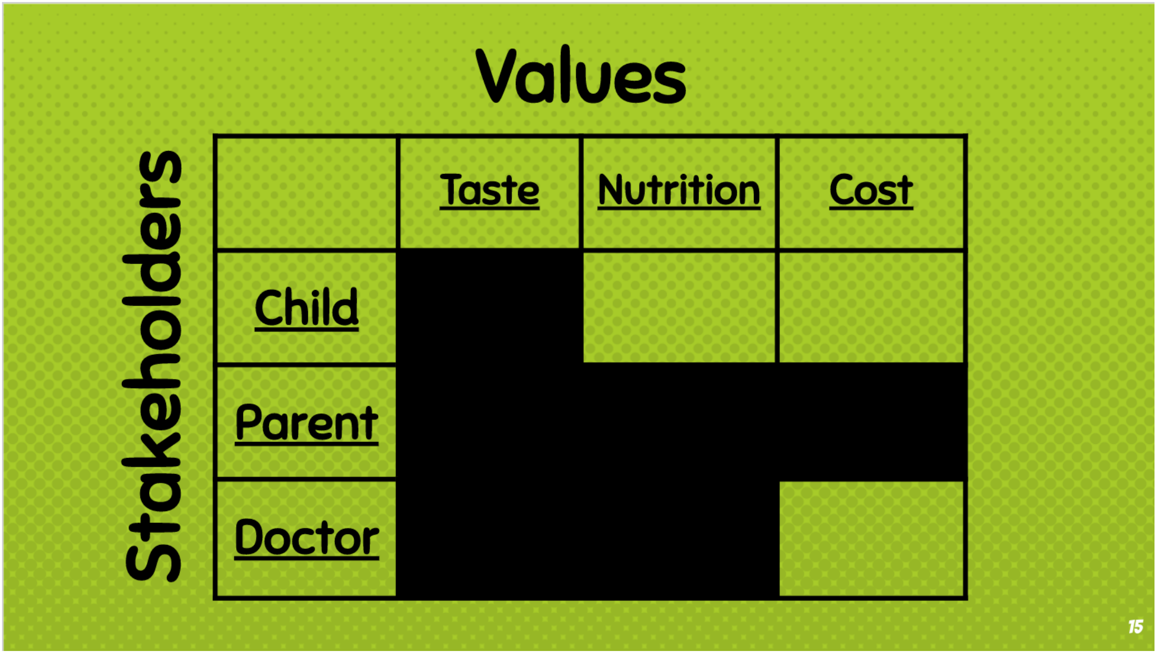
The “best” PB&J sandwich could mean a lot

8

Figure 8-3: As a class, students discuss that the “best” PB&J sandwich could mean the most delicious sandwich, the healthiest, or the quickest to make.

This activity was included to show students how opinions can be “baked” into an algorithm, even one that appears to have an apparent goal. This activity was also included because it is common in U.S. K-12 computer science classrooms to introduce algorithms in comparison to a recipe and to have students write an algorithm for making a peanut butter and jelly sandwich. Usually the focus of this exercise is on the need to write specific, unambiguous instructions, but this exercise adds, without much extra effort to the original activity, the emphasis that even in the specific operationalizing the construction of a PB&J sandwich, the goal for the algorithm can be ambiguous and subjective.

8.4.3 Ethical Matrix



Values

Stakeholders

	Taste	Nutrition	Cost
Child	[redacted]		
Parent	[redacted]	[redacted]	[redacted]
Doctor	[redacted]	[redacted]	

15

Figure 8-4: An example of an ethical matrix the class completes together before students.

Continuing from the previous lesson, students are then asked how they should decide which the goal their algorithm will "optimize." Similar to the pilot study, students are then presented with the concept of an ethical matrix. The instructor walks through an example of an ethical matrix for the peanut butter and jelly sandwich algorithm, with students driving most of the discussion for the last row of the ethical matrix. As a class, students then discuss how to use the matrix to decide what the goal of their algorithm should be. The ethical matrix could be used to see which value most stakeholders had in common, which could then translate to that value becoming a reasonable goal for the algorithm. The ethical matrix could also be used less prescriptively and more as a conversation starter to realize another goal for which the algorithm should optimize.

In small groups of three or four students, students then construct their own ethical matrices around the peanut butter and jelly sandwich algorithm. Before constructing the matrices, each group is asked to brainstorm ten direct and indirect stakeholders for their algorithm, and ten possible values various stakeholders might have for the PB&J

algorithm. They then fill out ethical matrices of varying dimensions (e.g., varied by the number of stakeholders or values they are required to consider). After constructing each matrix, students discuss in their groups what the goal of the algorithm should be and discover that varying the number of stakeholders and values, along with varying which stakeholders and values are present in the matrix, affect which goal is decided upon.

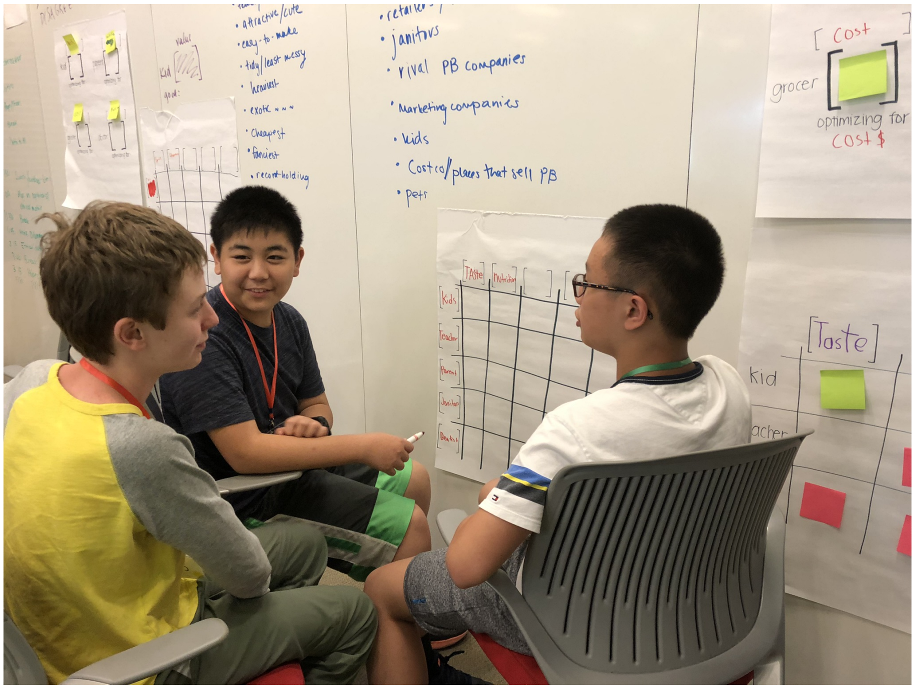


Figure 8-5: A group of students work on filling out various ethical matrices for their PB&J sandwich algorithm.

8.4.4 Introduction to Supervised Machine Learning and Algorithmic Bias

This activity is very similar to Activity #1 presented in the Pittsburgh pilot. It begins by introducing students to the concept of supervised machine learning and classification. Supervised machine learning is presented in comparison to a baby learning her colors - often by having a parent show the baby many examples of something blue and saying "blue" and showing many examples of something red and saying "red." Classification is introduced through examples with varying subtlety. The first example provided to students is that of email spam filters, where it is clear that the algorithm is trying to classify where to place an incoming email. Less obvious examples like face detection (where the two classes are: image includes a face or image does not include a face) and handwriting detection (where the classes are the numerical digits 0-9) are then presented.

After this class discussion, students are given time to play in partners with Google's 2017 (version 1) of Teachable Machines. Teachable Machines is a web browser tool that allows users to build their own image or sound classification models through a visual interface. That is to say, no programming is required. There are two versions of Teachable Machines: the 2017 version (version 1), which is intended as a visual explainer for supervised machine learning and classification, and the 2019 version (version 2), which is more robust and can be used as a both a teaching tool and a scientific one.

After this exploratory period, students are asked to share what they have been trying to classify and some general patterns they have found when changing their training dataset. Students are then tasked, in pairs, to do the same activity presented in the Pittsburgh pilot with the 2019 version (version 2) of Teachable Machines. They are asked to build a cat-dog image classifier but are unknowingly given a biased dataset where the images of cats are overrepresented and more diverse than the images of dogs. After training their models on these images, students record the accuracy of their classifiers. When the classifier is more accurate for the images of cats than the images of dogs, students are given additional images to retrain their classifiers with their own new datasets and record the accuracy of their new models.

Similar to the pilot, this activity ends with students watching Joy Buolamwini's YouTube video entitled "Gender Shades" and are asked to connect their findings from the activity to Buolamwini's research findings [35]. That is, students connect the algorithmic bias they have found in their own models to the bias Buolawmini has found in commercial facial recognition systems. They then discuss how technology companies could make facial recognition systems less biased against darker-skinned women by changing their training datasets.

Figure 8-6: A student experiments with training his own supervised machine learning model with Google's Teachable Machine platform.

8.4.5 Speculative Fiction

This activity shifts away from algorithmic bias to other ethical issues that arise with technology. In this activity, students have the opportunity to interact with various AI technologies, such as emotion detection software, AI-enabled image editing and image producing software, or AI-enabled text generators, all listed in Appendix C. Students then respond to creative writing prompts about who might be affected by the technology and how the technology might produce harm or benefit in the future. They turn their responses into slides and present their speculative envisionings to their classmates, and discuss as a class.

This activity, based on several activities employed at the undergraduate and graduate level, was added to get students to think about ethical issues raised by artificial intelligence beyond the issue of algorithmic bias [29, 38, 51]. This activity aims to get students to consider the direct, secondary, and tertiary consequences of creating an AI system. The forecasting abilities practiced here give students the opportunity to anticipate otherwise unanticipated consequences of embedding AI systems into existing social structures; that is, this activity helps students to further see AI systems as socio-technical systems.

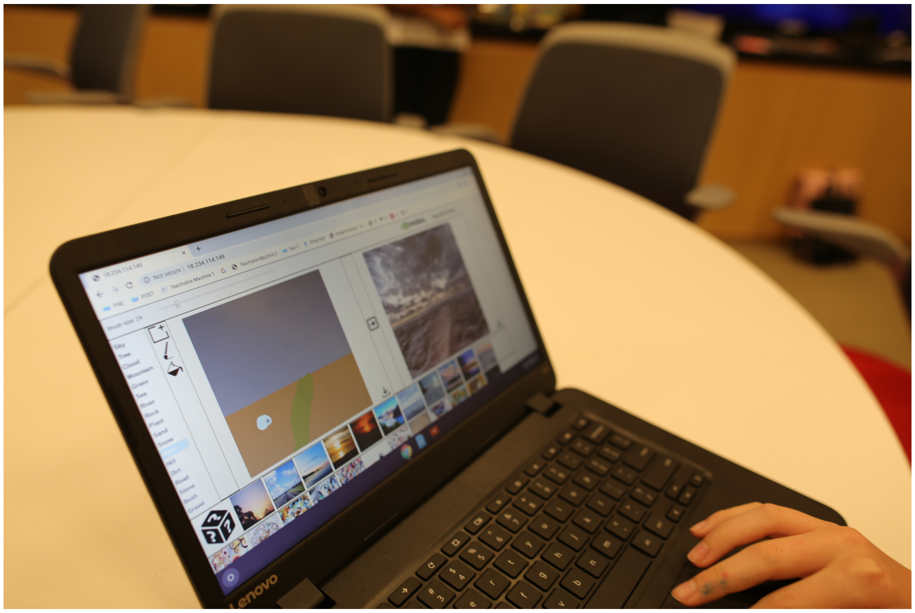


Figure 8-7: A student interacts with a GAN tool before engaging in the speculative ficition activity.

8.4.6 YouTube Scavenger Hunt

This activity is similar to the previous AI Bingo activity. In partners, students are asked to recognize the various AI systems on the YouTube platform (e.g., advertisement matching algorithm, the recommender algorithm, comment classifier, etc.). For each system, students identify what the algorithm is trying to predict and the dataset the algorithm uses. This activity was included in order to get students thinking about the many different ways AI is used within a single platform and to prepare them for the capstone YouTube Redesign activity.

8.4.7 YouTube Redesign Activity

Similar to the third activity described in the Pittsburgh pilot, students apply what they have learned so far to the many AI systems related to YouTube, with a special emphasis on their recommendation algorithm. Students are asked to identify ten

Figure 8-8: Students work on their paper prototypes for the YouTube Redesign activity.

stakeholders in YouTube, ten values that stakeholders in YouTube might hold, and then are asked to construct an ethical matrix for the platform. Based on this ethical matrix, students determine a goal (or "opinion") for their algorithm.

Students are then introduced to a variety of "redesigned" social media examples, such as examples from designer Tristan Harris's website, Gobo.social, and are asked to discuss how the different design choices lend themselves to different goals as well as enable different kinds of data collection [62, 15]. Students then are given craft materials to paper prototype what this new version of YouTube would look like based on their ethical matrix and are asked to imagine features that meet the values their identified stakeholders have.

8.4.8 YouTube Socratic Seminar

Lastly, as a concluding activity, students read an abridged version of a Wall Street Journal article titled "YouTube Weighs Major Changes to Kids Content Amid FTC

Probe." The article touches upon YouTube's lack of COPPA compliance and the changes that the company is considering in order to make the platform safer for young viewers. After reading the article together, students participate in a socratic seminar discussing which stakeholders are most important or influential to the proposed changes to the YouTube Kids app and whether or not technologies like autoplay should exist. Students are asked questions about the article such as:

- Can you summarize this article?
- Which stakeholders were mentioned?
- Do you think YouTube kids should be a separate product? Why? Why not?
- How do you think advertisers feel about these changes?
- Do you think YouTube will lose profit if they decide to make the changes? Is it okay if they lose profit?

All questions can be found in the larger curriculum document. During the conversation, students are encouraged to respond to other's comments and pose their own questions.

"It is not only possible that everybody
get involved with AI, it is actually
crucial that you do so."

Rachel Thomas

9 Study Protocol

In order to evaluate the curriculum's effectiveness in teaching students the concepts outlined in the standards, the following study was performed.

9.1 Format

These activities were piloted during a five-day summer workshop. The summer workshop format was selected based on the design recommendations of the previous pilot, which indicated that students could use longer periods of time to work on some of the activities. The workshop was open to students entering grades 5 - 9 in the Boston area and was held in classrooms on MIT's campus. Students were split into two separate classes depending on grade. Class A was composed of students entering grades 5-7, while students entering grades 7-9 were in Class B (seventh graders were ran-

Figure 9-1: Some of the workshop students having fun on a tour of MIT.

domly assigned to either class). Throughout the week, students engaged in all eight activities in the order that they are presented above, as well as other activities such as campus tours, outdoor time, and arts and crafts, as seen in Fig. 9-1.

9.2 Instructors

The instructors for this workshop included two graduate students, one undergraduate student, and three instructors from local after-school STEM programs. The two graduate students acted as the main instructors, providing short lectures and facilitating class discussions, while the others acted as teaching assistants. Teaching assistants were given all of the materials ahead of time and completed a one-day training before the camp to understand the goals of the curriculum. At the end of each day, all staff debriefed the activities that were done that day by discussing what worked well, what didn't work well, and what they might change in the future.

9.3 Participants

We recruited students entering grades 5-9 in the Boston area through a local for-profit STEM education program, Empow Studios, as part of their summer workshop series. They were recruited through flyers and online postings by Empow Studios. It is important to note that this group of students self-selected into the workshop, and are likely not representative of all middle school aged students. Parents gave written informed consent and students gave signed assent prior to the study. All students were exposed to the same curriculum over the course of five days. The protocol was approved by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES).

Ultimately, a total of 28 students participated in the study, with 21 completing all five days of the workshop. Students had a wide range of computer science experience, ranging from none at all to more than seven years. Additional demographic information is available in Table 9.1.

It is worth noting that 19 students were ultimately used to do the analysis on the survey data presented in the following section. This is due to one student who was absent during the pre-workshop survey and two students failing to pass an attention check in the post-workshop survey.

9.4 Data Collection

Several modes of data collection were used throughout the week, each discussed in more depth below. Each means of data collection was designed to provide insight on one or more of the following research questions:

1. What are children's preconceptions of artificial intelligence and the ethical issues associated with it?

2. What are children capable of learning about key concepts in AI and ethics? How does this inform a developmentally appropriate curriculum?

3. How does learning about the ethical design of artificial intelligence change children's perceptions of artificial intelligence?

4. How does learning about the ethical design of artificial intelligence change their perceptions of themselves as empowered designers?

9.4.1 Workshop Surveys

Students were administered a pre-workshop survey and a post-workshop survey, which can be found in Appendix B. The survey questions can be categorized into one of three groups. First, questions 1 and 46 sought to understand which technologies students engage and their ability to recognize which of those technologies are examples

Demographics	**Breakdown**	
	Category	#
Grade	Fifth	8
	Sixth	5
	Seventh	6
	Eighth	7
	Ninth	2
Gender	Male	20
	Female	8
Heard of AI before workshop?	Yes	23
	No	5
Role of Technology	Can't live without it	25
	It scares me	2
Days Participated	Monday	28
	Tuesday	27
	Wednesday	27
	Thursday	22
	Friday	21

Table 9.1: Participant Demographics

of artificial intelligence. Students were asked to list which technologies they used regularly and to label which of those they thought involved AI. Second, questions 3-39 sought to understand what narratives about AI students were already familiar with, their understanding of how it works, and their perceived role in the design of AI. These questions were adapted from Cave, Coughlan, and Dihal (2019) and asked participants if they were familiar with a particular narrative surrounding AI (for example, the idea that AI will help with everyday tasks or will displace workers), how likely they thought that narrative is, and how concerned or excited about each narrative they were [40]. This set of questions also asked students to define AI in their own words and answer questions about their ability to shape the future of AI. Question 43 also asked about students' preconceived notions of AI by asking students how much they thought a series of words described AI (such as objective, subjective, trustworthy, helpful, etc). Finally, questions 40, 41, and 42 addressed students' ability to identify who was affected by artificial intelligence by asking them who is affected by technologies like YouTube, self-driving cars, and Google Search.

9.4.2 Mural and Interviews

Throughout the week students also participated in creative reflections and interviews. Students ended each day in the workshop by contributing to a class mural (with the exception of the first day when they participated at the very beginning of the workshop and at the end of the day). Students were given different shapes representing different reflective tasks and asked to add two shapes to the mural. Students could add a question, hope, concern, feeling, or drawing about AI to the mural. Each day's shapes were color coded to the day.

Additionally, students engaged in exit slips or recorded interviews at the end of each day. Students were asked what they added to the mural, what they learned that day, which activity they learned the most from that day, and which activity was their favorite.

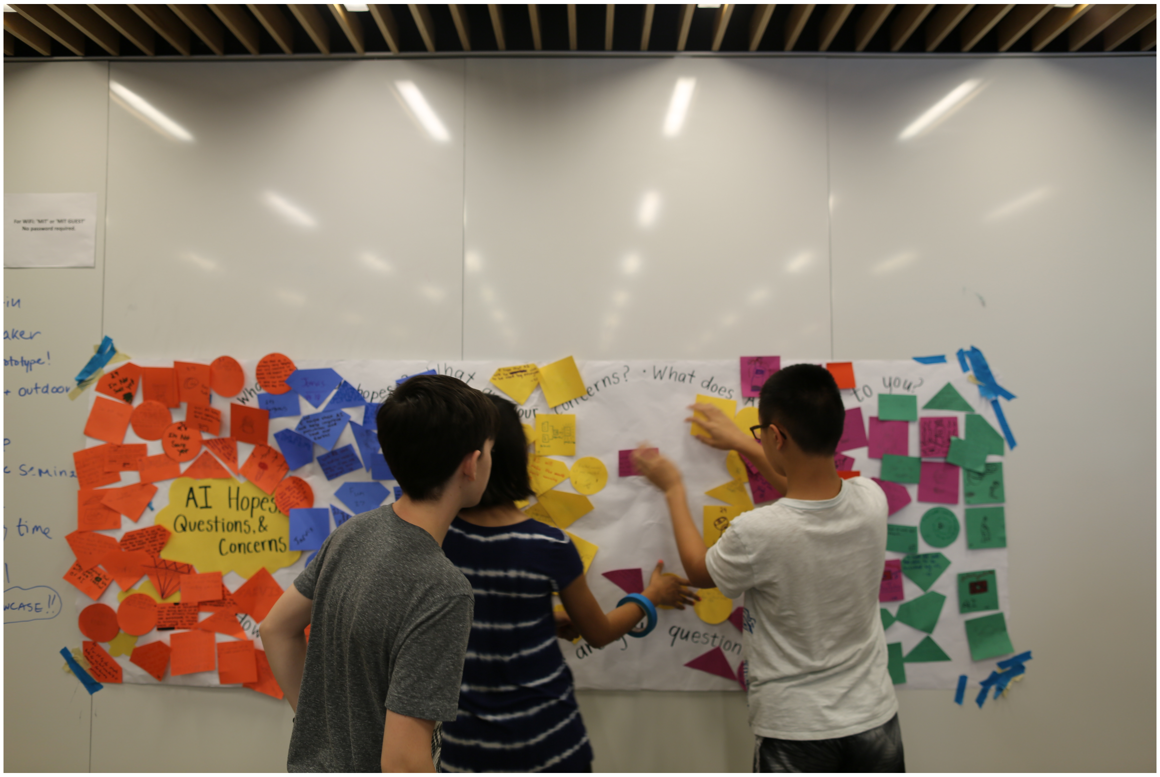


Figure 9-2: Students add their hopes, questions, concerns, feelings, and drawings of AI to the classroom mural.

9.4.3 Data Collection for Activities

Finally, all of the students' work was the last source of data for analysis. Students' AI Bingo cards, slide decks from the Speculative Fiction exercise, and their YouTube Scavenger Hunt activity sheet were all kept and recorded. Their YouTube Redesign projects and the corresponding activity sheet were kept as artifacts for analysis as well. The Socratic seminar was recorded and transcribed. The only activity which had its own post-assessment was the algorithmic bias activity, where students were asked a few questions to show their understanding of how a supervised machine learning algorithm works and their understanding of how algorithmic bias is shaped by training datasets.

"This is the future I'm hoping for.... One where we stop seeing machines as objective masters and start treating them as we would any other source of power."

Hannah Fry

10

Results

In this chapter, I present data and analysis to the research questions posed in the previous chapter.

10.1 What Do Children Already Know about AI?

To answer the question about what middle schoolers already know about AI, students were asked to complete the pre-workshop survey presented in Appendix C. Questions 3-39 were adapted from [40], where the authors surveyed over one thousand adults in the United Kingdom about their perceptions of AI. In their work, Cave, Coughlan, and Dihal found that 85.0% of respondents had heard of AI while 11.0% said they had not, and 4.0% said they were unsure. Among the students, 80.7% said they had heard of AI, while 19.2% had not. In their paper, Cave et al. report that 42.0%

Technology	Breakdown	
	Subtype	#
YouTube		18
Google Search		18
Hardware		12
	Gaming System	6
	Tablet	6
Voice Agents		12
	Siri	8
	Alexa	5
	Google Home	4
	Cortana	2
Netflix		11
Email		11
Amazon		10
Music Streaming Service		9
	Spotify	7
	Pandora	3
Other Social Media		7
	Instagram	4
	Other	4
	Discord	2
	Amino	1
	Reddit	1
	TikTok	1
	Twitter	1
Computational Thinking Tool		6
	Scratch	5
	Coding	2
	KTByte	1
	Lego Mindstorms	1
	Python	1
Google Hangouts		6
Other		4
Texting		2

Table 10.1: The number of participants that use each technology. There were 19 participants in total.

of participants describe AI relative to human cognition, using phrases like "decision-making," "learning," or "thinking." In this workshop, 38.8% of the students described AI in these terms. One participant wrote, "It acts like a human. It can write poems, write music, and do other things. It is a technology version of a human." Another participant wrote succinctly, "People but robots."

On the topic of robots, 38.8% of the students mentioned "robots," whereas 25.0% of the participants in the study by Cave et al. mentioned robots. The authors also found that 12.0% of their participants incorporated hopes or fears into their descriptions of AI. Similarly, 11.1% of the participants included similar depictions in their answers. One participant wrote, "I would describe it as something that is very useful but can go very wrong. When you need something and you can't find it AI will help. But AI can take over people's jobs and if done wrong it can be very dangerous."

Technology	Does it use AI?	Student Response "True" (%)	Student Response "False" (%)	Student Response "Don't Know" (%)
Google Search	Yes	**84.2**	10.5	5.2
Wireless Printer	No	26.3	**52.6**	21.0
FaceTime	No	31.5	**52.6**	21.0
Nintendo Switch	No	52.6	**21.0**	26.3
Instagram Feed	Yes	**63.1**	10.5	26.3
YouTube Subscriptions	No	52.6	**31.5**	15.7
Netflix Recommendations	Yes	**89.4**	5.2	5.2
Snapchat Filter	Yes	**63.2**	21.0	15.7
GPS	Yes	**73.6**	15.7	10.5
Alexa Reminders	No	57.8	**21.0**	21.0

Table 10.2: Students' initial conceptions about which popular technologies use AI.

It is also worth noting that in the students' descriptions of AI, 27.7% used words like "programming" or "coding," implying that they recognize that AI is built with software. Similarly, 16.6% mentioned the usage of data or information as essential to producing AI.

In addition to the questions posed by Cave et al., some questions were added to the pre-workshop survey to see what technologies children use and if they are able to recognize examples of AI in the world around them. Table 10.1 displays the technologies children reported using regularly.

With respect to identifying AI in the real world, children were asked to state whether or not they believed the following technologies used AI: Google Search, Wireless Printer, FaceTime, Nintendo Switch, Instagram Feed, YouTube Subscriptions, Netflix Recommendations, Snapchat Filter, GPS, Alexa Reminders. Table 10.2 summarizes their responses.

10.1.1 What Do Children Already Perceive about the Ethics of AI?

Not only is it necessary to know what children know about the basic functions of AI, it is also necessary to understand how they see it situated in society. Continuing to build on the work by Cave et al., students were surveyed about their general perception of technology. Students were asked which of the following sentiments best represented the role technology plays in their life: "it scares me" or "I can't live without it." Only one child out of eighteen chose "it scares me," while Cave et al. reported that 1% of UK adults also reported "it scares me." Then, students completed a survey to find out their familiarity with eight common narratives surrounding AI (as originally determined by Cave et al.). These narratives include four hopes: ease, dominance, immortality, and gratification. They are:

- Immortality refers to how AI might be used in pursuit of health, for example through personalized medicine.
- Ease refers to the hope that AI will increasingly perform many tasks that people do not want to do.
- Gratification refers to the way one might wish to use that free time made by AI assuming it performs the tasks described in Ease.

- Finally, Dominance, or power over others, can be seen as the means to protect this blissful existence through AI contributing to powerful new means of defense and security.

For each hope there is also a paralleling narrative around fear, and the authors list the following four:

- Inhumanity: refers to the risk that in the pursuit of an ever longer lifespan, a person loses their identity, becoming more machine than human.
- Obsolescence refers to the fear of being put out of work.
- Alienation refers to the fear that humans become alienated from each other and prefer to interact with machines.
- Uprising refers to the fear that AI-enabled power will turn on them.

Similar to Cave et al., obsolescence and ease were the narratives surrounding AI that children were most familiar with, followed by uprising, dominance, and immortality. Only one child had not heard of any of these narratives around AI. Table 10.3 shows how many students recognized each phrase, and how they compared to the adult population surveyed by Cave et al.

Furthermore, students were asked how they felt about each narrative on a scale of concern to excitement, how likely they thought each narrative would come true, and how likely it might come true in their lifetime. Students were asked to rate how excited or concerned they felt about each narrative on a scale of 1-5, where 1 equalled concerned and 5 equalled excited. Similar to Cave et al., scores 1-2 are counted towards the total percentage "concerned," scores 4-5 are counted towards "excited." Students were most concerned about the narratives of uprising and obsolescence, which are also the narratives that UK adults were most concerned about. Students were most excited about the immortality narrative, which the UK adults surveyed were also most excited about. Table 10.4 reports the results.

Students were also surveyed how likely they felt these narratives were to ever occur. Here, scores 1-2 are counted towards the total percentage "unlikely" and scores 4-5 are counted towards "likely."

Narrative	Recognition by UK Adults (%)	Recognition by Middle School Students (%)
Obsolescence	55	**88.8**
Ease	53	**88.8**
Uprising	44	**77.7**
Dominance	30	**72.2**
Alienation	20	**66.6**
Immortality	19	**72.2**
Gratification	16	**66.6**
Inhumanity	13	**72.2**
None of the Above	6	**5.5**

Table 10.3: Students' familiarity with eight common narratives around AI.

Narrative	Excitement (%)	Concern (%)	Average Score
Ease	50.0	22.2	3.44
Immortality	50.0	27.7	3.27
Gratification	38.8	27.7	3.16
Dominance	27.7	27.7	3.11
Inhumanity	16.6	66.6	2.22
Alienation	11.1	55.5	2.33
Obsolescence	11.1	77.7	2.00
Uprising	0.0	66.6	1.88

Table 10.4: Students rate how excited or concerned they are about different narratives surrounding AI.

Narrative	Likely UK Adults (%)	Likely Middle Schoolers (%)	Unlikely UK Adults (%)	Unlikely Middle Schoolers (%)
Ease	48	**73.6**	5	**0.0**
Dominance	42	**73.6**	7	**0.0**
Alienation	18	**52.6**	25	**21.0**
Gratification	18	**47.3**	26	**15.7**
Obsolescence	35	**47.3**	12	**10.5**
Inhumanity	12	**36.8**	30	**26.3**
Uprising	30	**36.8**	16	**26.3**
Immortality	19	**15.7**	28	**42.1**

Table 10.5: Students rate how likely or unlikely they believe different AI narratives are to come true.

Finally, students responded if they perceived these narratives would impact them in their lifetime. Again, scores 1-2 are counted towards the total percentage "unlikely to impact in my lifetime," scores 4-5 are counted towards "likely to impact in my lifetime."

These findings were corroborated by students' additions to the mural at the beginning of the week, before instruction began. The most frequent themes from students' addition to the mural at this point in time were of ease, obsolescence, and uprising. One student wrote, "I hope AI will assist in helping future beings," while another wrote, "I feel that AI is helping in day to day life, and that it can help us farther in the future." One student elaborated, "I feel that AI will help us a lot, but it also poses a potential problem. AI will help us save time, and do less work. AI could also make bad decisions, or 'takeover' the world. AI is good, in my opinion."

In addition to asking students about their hopes and fears around AI, students were also asked to rate how strongly they agreed with several descriptions of AI: good, bad, complicated, simple, easy to understand, magical, dependable, factual, trustworthy, helpful, fair, biased, unbiased, objective, and subjective. Of these descriptions,

Narrative	Impact in my lifetime UK Adults (ages 16-34) (%)	Impact in my lifetime Middle Schoolers (%)
Ease	53	**68.4**
Dominance	34	**57.8**
Obsolescence	33	**57.8**
Alienation	12	**52.6**
Uprising	20	**42.1**
Gratification	12	**36.8**
Inhumanity	7	**31.5**
Immortality	14	**26.3**

Table 10.6: Students rate how likely or unlikely they believe different AI narratives will have an impact during their lifetime.

students reported that the term "helpful" was the best description of AI as it had the highest average score. This is consistent with the previous finding that one of the most popular narrative students recognized about AI was the "ease" narrative. After "helpful" descriptions like "complicated," "factual," "good," and "objective" had the next highest scores. The descriptions with the lowest scores are "magical" and "simple."

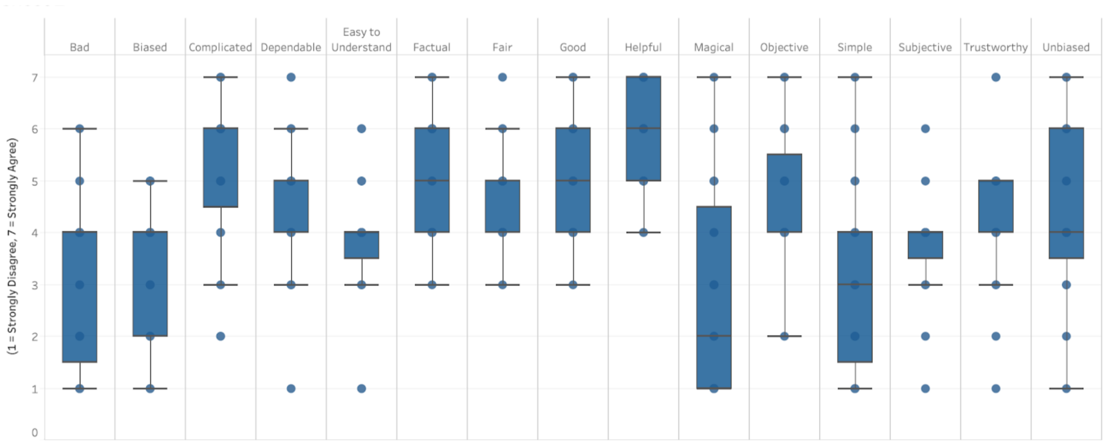


Figure 10-1: How strongly students agree or disagree with various descriptions of artificial intelligence.

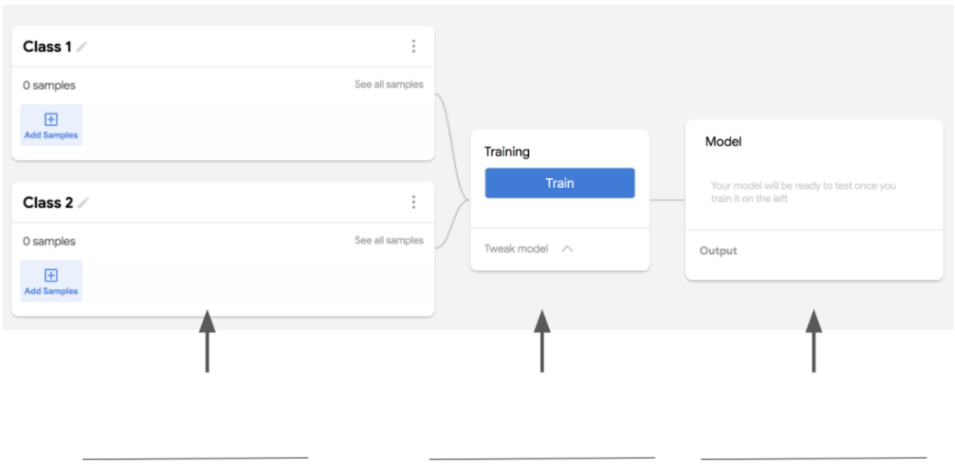
10.2 What Can Children Learn?

After examining students' preconceived notions about AI and ethics, we must explore what students learned after the pilot intervention. The results below are ordered according to the suggested standards put forth in Chapter 8.

10.2.1 Understand the Basic Mechanics of Artificial Intelligence Systems

The first set of suggested standards centers on students' understanding of AI systems: being able to recognize them in the real world, being able to broadly label the components of an AI system (as a dataset, learning algorithm, and prediction), and being able to understand how classification works in the supervised machine learning context. Students' understanding of these concepts were assessed after the Teachable Machines activity.

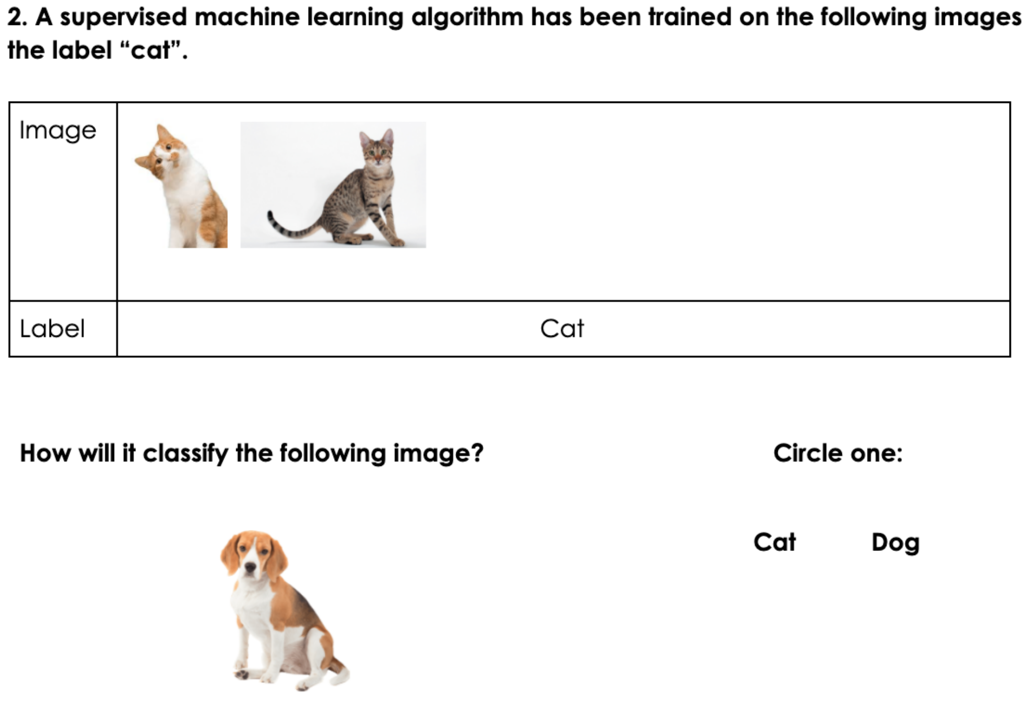
1. What are the three components of an AI system?



Your model will be ready to test once you train it on the left

Figure 10-2: Assessment used to check students' understanding of the three parts of a supervised machine learning system.

Students were asked to label the three components of an AI system, as seen in Figure 10-2. Recall that the definition given to them in the context of supervised machine learning was a dataset, learning algorithm, and a prediction. When asked, 11.1% of students labeled all of the components correctly, and 77.7% of the students labeled two of the three components correctly. The most common mistakes were mislabeling "learning algorithm" as simply "algorithm." When asked, 22.2% of students labeled it correctly, and of those who labeled it incorrectly, 55.5% of students made this mistake. When asked, 61.1% of students correctly labeled "dataset," but 27.7% of students who mislabeled it, labeled it as "input." Finally, 72.2% of students knew the last piece of the system was a "prediction," but of the students who mislabeled this, 27.7% of students labeled it as "output" instead.



2. A supervised machine learning algorithm has been trained on the following images with the label "cat".

Image	
Label	Cat

How will it classify the following image?

Circle one:

Cat **Dog**

Figure 10-3: Assessment used to check students' understanding of how supervised machine learning works.

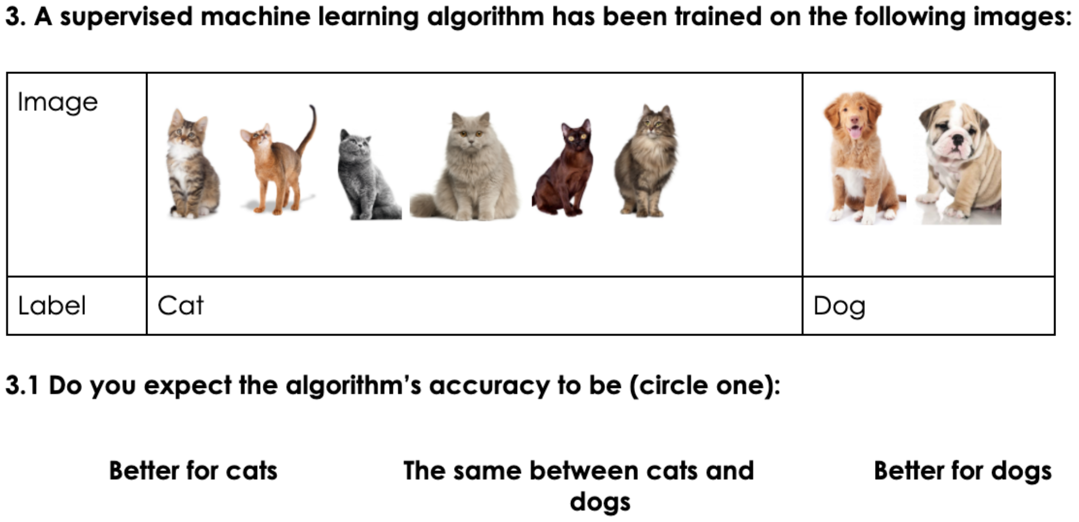
Students also answered questions related to supervised machine learning, classification, and algorithmic bias. The first question was adapted from [104] as seen in Figure 10-3. Students were told that a supervised machine learning algorithm had been trained on two images of cats. When presented with an image of a dog, they were asked to predict how the algorithm would classify the image (the correct answer

Technology	Does it use AI?	% Student Correct (Pre)	% Student Correct (Post)
Google Search	Yes	84.2	94.7
Wireless Printer	No	52.6	57.8
FaceTime	No	52.6	57.8
Nintendo Switch	No	21.0	21.0
Instagram Feed	Yes	63.1	89.4
YouTube Subscriptions	No	31.5	21.0
Netflix Recommendations	Yes	89.4	94.7
Snapchat Filter	Yes	63.2	78.9
GPS	Yes	73.6	78.9
Alexa Reminders	No	21.0	15.7

Table 10.7: Students' conceptions about which popular technologies use artificial intelligence after learning how supervised machine learning works.

being cat, since the algorithm had been given no examples of the "dog" class). When asked, 77.7% of students answered this question correctly.

Students' ability to recognize AI systems in everyday technologies was assessed again in the post-workshop survey. The results can be seen in Table 10.7.



3. A supervised machine learning algorithm has been trained on the following images:

Image		
Label	Cat	Dog

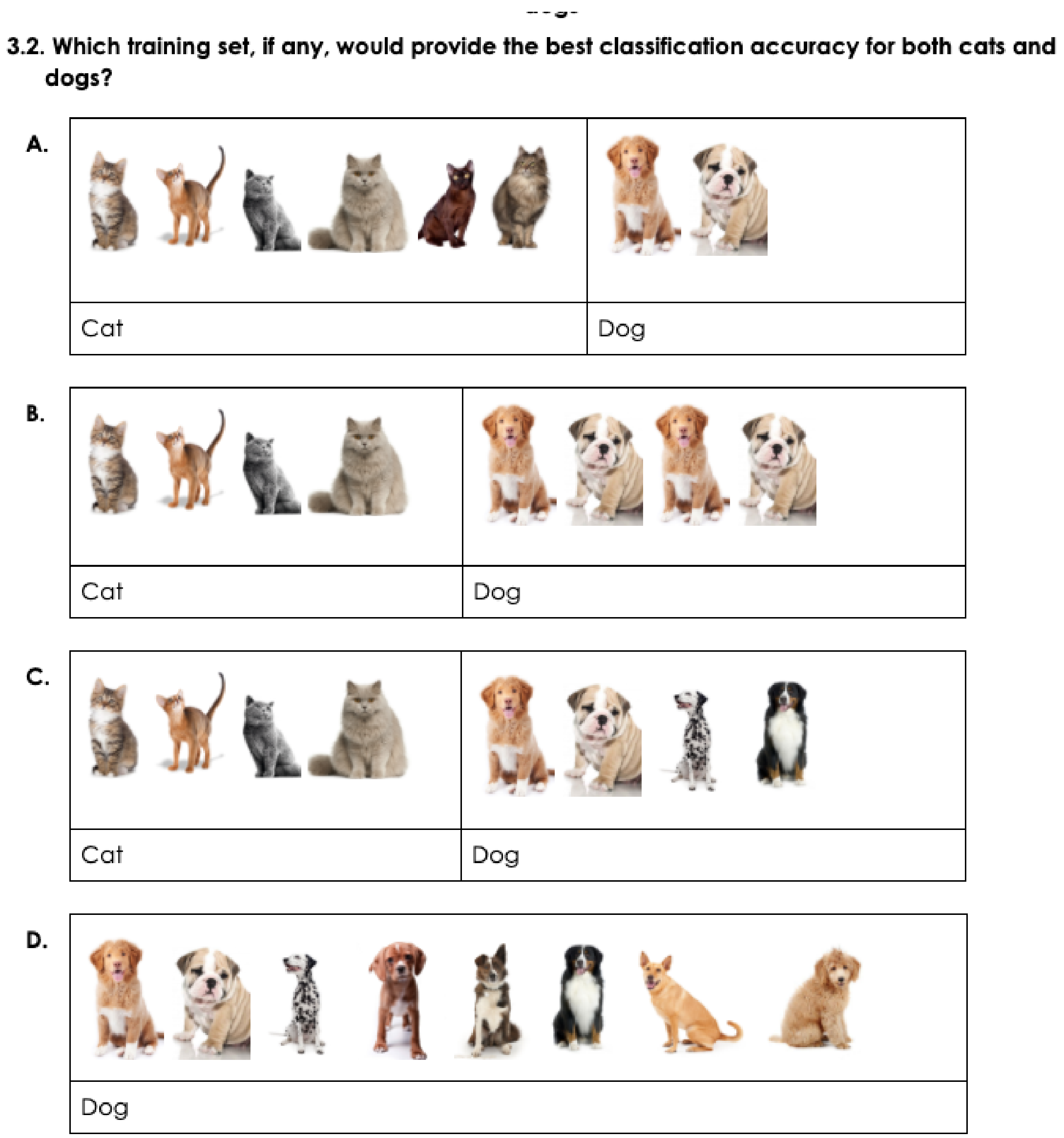
3.1 Do you expect the algorithm's accuracy to be (circle one):

Better for cats **The same between cats and dogs** **Better for dogs**

Figure 10-4: Assessment used to check students' understanding of the effect training data has on the accuracy of a machine learning system.

Students were also assessed on their understanding of algorithmic bias after the corresponding activity. As seen in Figure 10-4, students were shown an image dataset consisting of six cats and two dogs and asked which group the algorithm would be more accurate with, if any. When asked, 100% of students correctly identified that the algorithm would be more accurate with images of cats than with images of dogs.

Students were also shown multiple datasets with varying representations of cats and dogs and were asked which dataset would most likely lead to the best classification accuracy on cats and dogs. This question can be seen in Figure 10-5. When asked, 100% of students also answered this question correctly, choosing the dataset where both cats and dogs were equally represented and where both groups had lots of diversity among the images.



dogs.

3.2. Which training set, if any, would provide the best classification accuracy for both cats and dogs?

A.

Cat	Dog

B.

Cat	Dog

C.

Cat	Dog

D.

Dog

Figure 10-5: Assessment used to check students' understanding of how the composition of training data affects the outcome of a supervised machine learning system.

10.2.2 Recognize the Many Stakeholders in a Socio-Technical System

In addition to students' technical abilities, many pieces of information give us insight into the development of their ethical design skills. One such piece of information is their ability to recognize that many stakeholders are involved in any technical system and that those stakeholders may value different aspects of the technology. In order to gauge students' ability to consider stakeholders, students were asked in the pre- and post-surveys three open ended questions:

1. What groups of people are concerned with the outcomes of YouTube's video recommender?
2. What groups of people are concerned with the outcomes of self-driving cars?
3. What groups of people are concerned with the outcomes of Google's search engine?

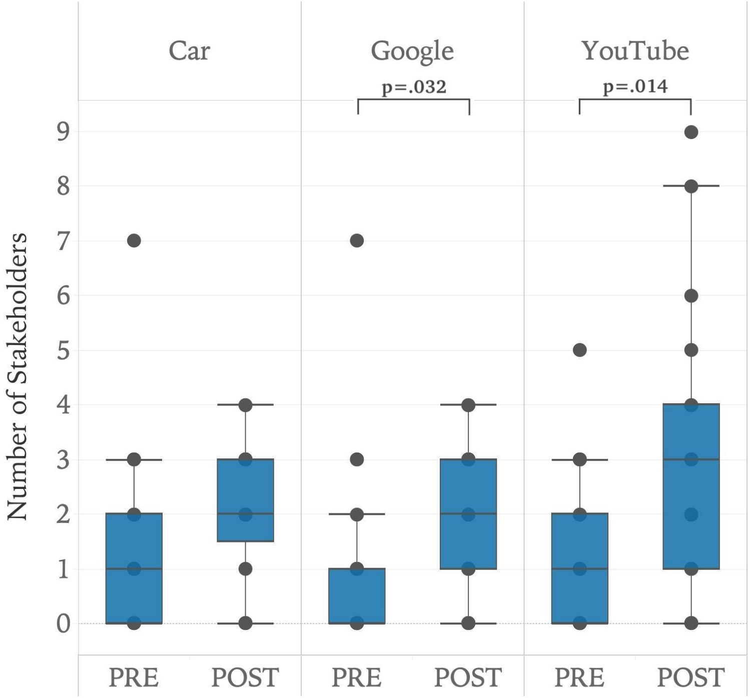


Figure 10-6: Assessment used to check students' understanding of how the composition of training data affects the outcome of a supervised machine learning system.

Each open ended answer was analyzed for the number of unique stakeholders identified. The average number of stakeholders from each class was calculated and compared from the beginning of the week to the end of the week using a Wilcoxon Signed Rank Test. The average number of stakeholders identified for YouTube Recommender System was higher at the end of the week (M=3.0, SD=2.56) than the beginning of the week (M=1.21, SD=1.44), with *p=.015*. Stakeholder identification for Google Search also was higher at the end of the week (M=1.84, SD=1.38) versus the beginning of the week (M=1.05, SD=1.75), with *p=.032*. There was no major difference between stakeholders identified for a self-driving car. These results are depicted in Figure 10-6.

Students' YouTube Redesign projects also show students' ability to map out stakeholders to values. In the activity, each student pair brainstormed ten stakeholders along with two values per stakeholder. For this analysis, values were grouped into larger themes. For example, "Good content" and "Entertainment" were grouped into the "Entertainment" category, and "Privacy" and "Safety" became the "Safety" category. Stakeholder-value pairs were recorded and summed across all students. The stakeholder-value pairs that guided the students' redesigns were YouTube-Money (9), Kids-Entertainment (7), YouTubers-Money (6), YouTube-Entertainment (6), and YouTubers-Entertainment (6), seen in Figure 10-7. These pairs determined the goals for the algorithm. The most common goals were Entertainment (6) and Profit (2).

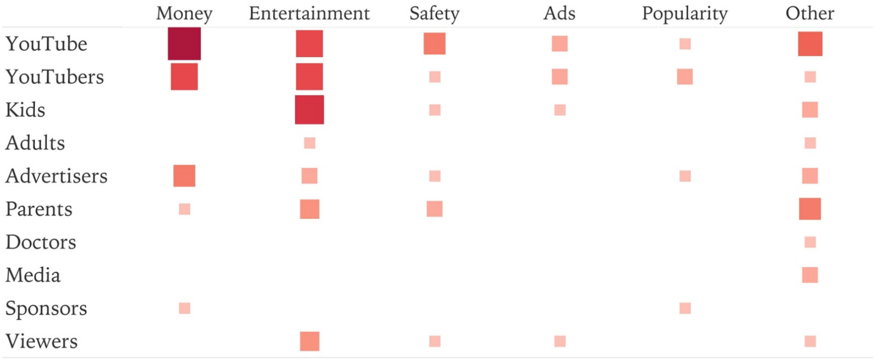


Figure 10-7: A visualization of the stakeholder-value pairs from students' YouTube Redesign projects.

Goal	Reasoning
Entertainment to be better	More people care about entertainment
Good Content	I think people like it more
Good content and fun	We decided on this goal because having fun is very fundamental. But being appropriate is also needed
Kid friendly and entertainment	They were the most important goals
Entertainment	Because the most stakeholders have the value of entertainment
Our goal is that our recommendation system will help with entertainment and money income	We decided on these two values, because overall, they were the most valued by the most people
Easily accessible, publicity	Out of the selected stakeholders, most say access and publicity are important
Profit	YouTube and YouTubers can use the profit to increase the quality of the site and videos
People get what they want. Also aware of how long they spend on the platform	If people get what they wanted they would stay on the site longer but not too long
To recommend popular videos, with few ads in multiple topics	Popularity is sought after most, ads are not as liked
N/A	Unable to provide justification

Table 10.8: YouTube Redesign Goals and Reasoning

Finally, the Socratic seminar discussion gives us additional context in understanding students' ability to identify stakeholders as they were asked throughout the discussion to justify their pairing of stakeholders to particular values. When they were asked who the most important stakeholders were in the FTC's investigation of YouTube, they identified several stakeholders similar to the ones that they addressed in their redesign projects. Children, parents, and Google were seen as the prominent stakeholders, and children were seen as the most vulnerable group out of the three. One student described why this was the case: “Because it's a lot about the safety of kids and what they watch because kids get easily influenced. So when [kids] see something's happening around them, they obviously think, ‘oh, they're more experienced we should copy whatever they're doing.’ So, it could be really bad that's why they take a long time to make sure everything's cautious and there's no bad content that could get released into the world of children.”

Another identified parents as a key stakeholder, because they were adults that had the children's best interests in mind: "I think parents are also a really important stakeholder because first of all kids do get influenced a lot, but still, parents are the ones who know what's appropriate and not appropriate, kids don't really know. So parents are kind of the ones that are more cautious and worried about the app being safe or not.”

One student pointed out that Google the company was the biggest stakeholder, because they had a financial interest in YouTube's success: "I think Google would be like one of the major ones because one thing, it's making lots of money on YouTube.”

10.2.3 Applying Their Understanding

Through the YouTube Redesign Project as well as the Socratic seminar, students had the opportunity to apply their understanding of AI systems, the importance of datasets, and the ability to design in light of many stakeholders with various and sometimes conflicting values.

Next, students were asked to reflect on the stakeholders and value pairs that they identified to come up with a goal for their redesign. These goals and the reasoning for them can be found in Table 10.8. The most common goals identified were Entertainment (5) and Good content (4). Seven out of the eleven pairs chose their goals because most stakeholders valued that particular characteristic. One pair chose their goal based on the perceived consequences of that goal (profit). Another pair chose their goal because they perceived it as morally superior to the other values in the table (Fun). One pair of students chose their goal because it seemed like a compromise among other values (aware of time spent on the platform). Lastly, one project was not able to come up with a goal or reasoning based on stakeholders.

Case Study 1: Optimizing for Good Content and Time Well Spent

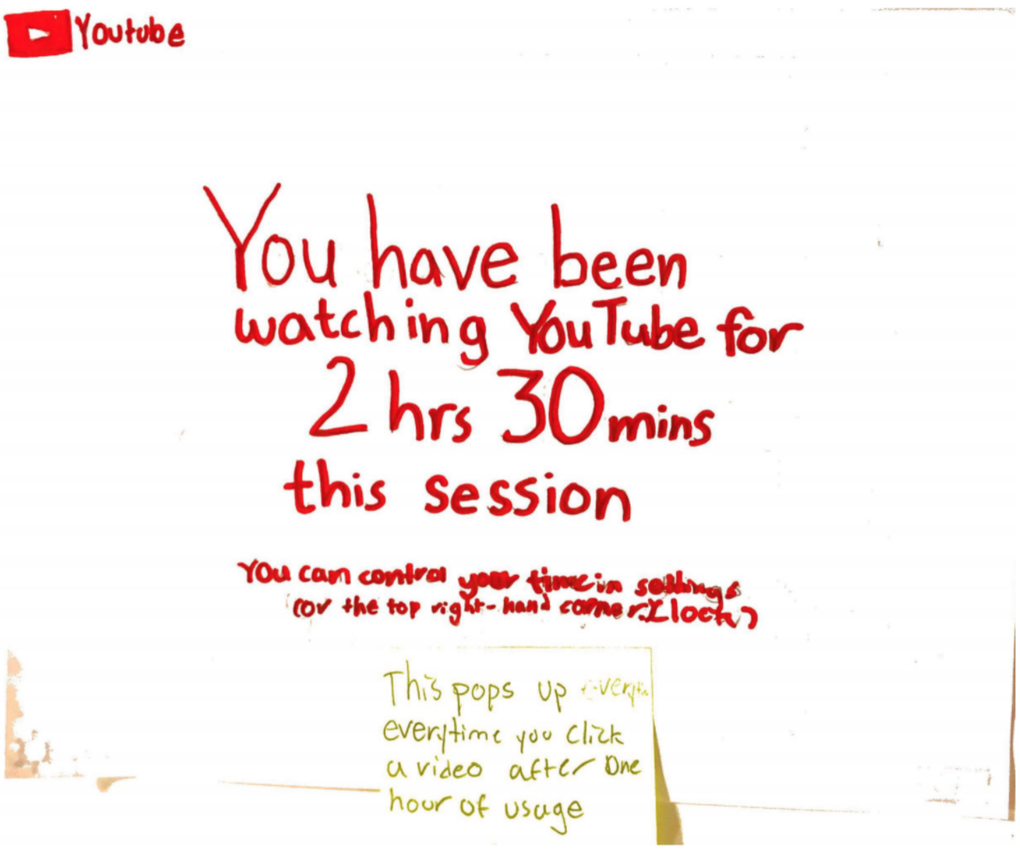

Youtube
You have been watching YouTube for 2 hrs 30mins this session
You can control your time in settings (or the top right-hand corner clock)
This pops up everyt
everytime you click a video after One hour of usage

Figure 10-8: A paper prototype of a YouTube optimizing for "time well spent."

This pair of a rising seventh and rising eighth grader decided that YouTube should maximize for “getting what they want and time well spent.” They identified YouTube, YouTubers, viewers, and sponsors as their stakeholders with the values of good content, popular, amount of ads, and money. Like previously mentioned, students chose this goal because "If people get what they wanted they would stay on the site longer, but not too long." Their paper prototype, seen in Figure 10-8, shows the one feature that they added to YouTube: a time tracker and pop up to let the user know how long they have been on the platform.

Case Study 2: Optimizing for Entertainment

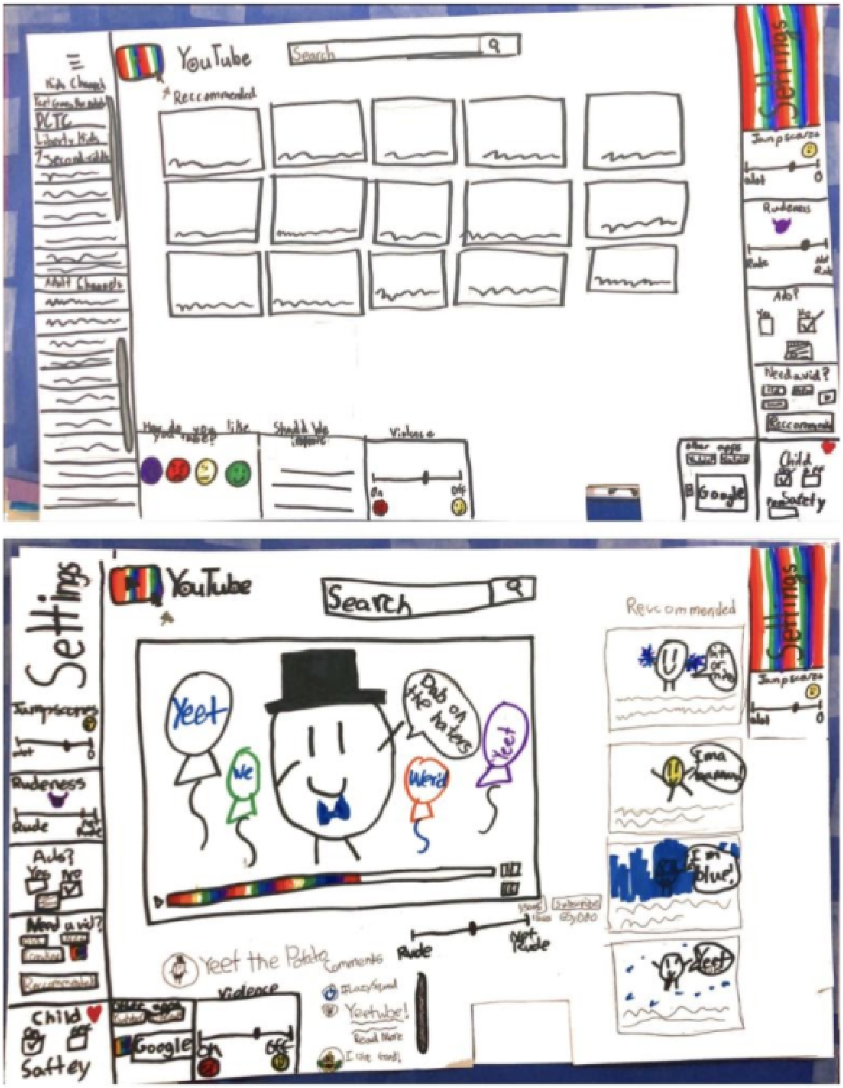


Figure 10-9: A paper prototype of a YouTube optimizing for entertainment.

This pair of rising fifth graders decided to optimize their new version of YouTube around entertainment. Students identified YouTube, kids, parents, and advertisers as their stakeholders along with the values of money, educational, good content, and fun, where most stakeholders were interested in "good content." The students wrote, "we decided on this goal because having fun is very fundamental but being appropriate is also needed." When asked how they would achieve this goal, the students wrote, "We would also teach it to give us clean content or videos that don't have swears. It will have a child safety mode to make sure there are no swears or inappropriate content." The home page of their design includes a section of videos for kids and a separate one for adults (Figure 10-9). The video page includes features such as a slider that allows users to choose levels of "rudeness" that appear in their videos and a child safety setting.

Case Study 3: Optimizing for Entertainment and Profit

This pair, including a rising fifth grader and a sixth grader, decided to optimize their new version of YouTube around both entertainment and profit. Students identified YouTube, viewers, YouTubers, advertisers as their stakeholders along with the values of entertainment, money, popularity, and safety, where most stakeholders were interested in "entertainment" and "profit." They explained the rationale behind the goal: "Popularity is sought out after most, ads are not as liked." These students focused on features that brought more income to YouTube, including filters for types of ads (leading to a more enjoyable ad experience by the user) as well as an option to donate directly to YouTube or YouTubers (Figure 10-10).

YouTube Redesign Trends

In addition to these projects, there are key trends to point out. The top five feature categories across all projects were: a feature related to ads, a child lock, a safe content feature, filters for certain video categories, and a feature for giving feedback. The most common feature was one related to ads, such as a button to turn off ads or a filter to make ads more relevant. Eight out of eleven projects included a feature

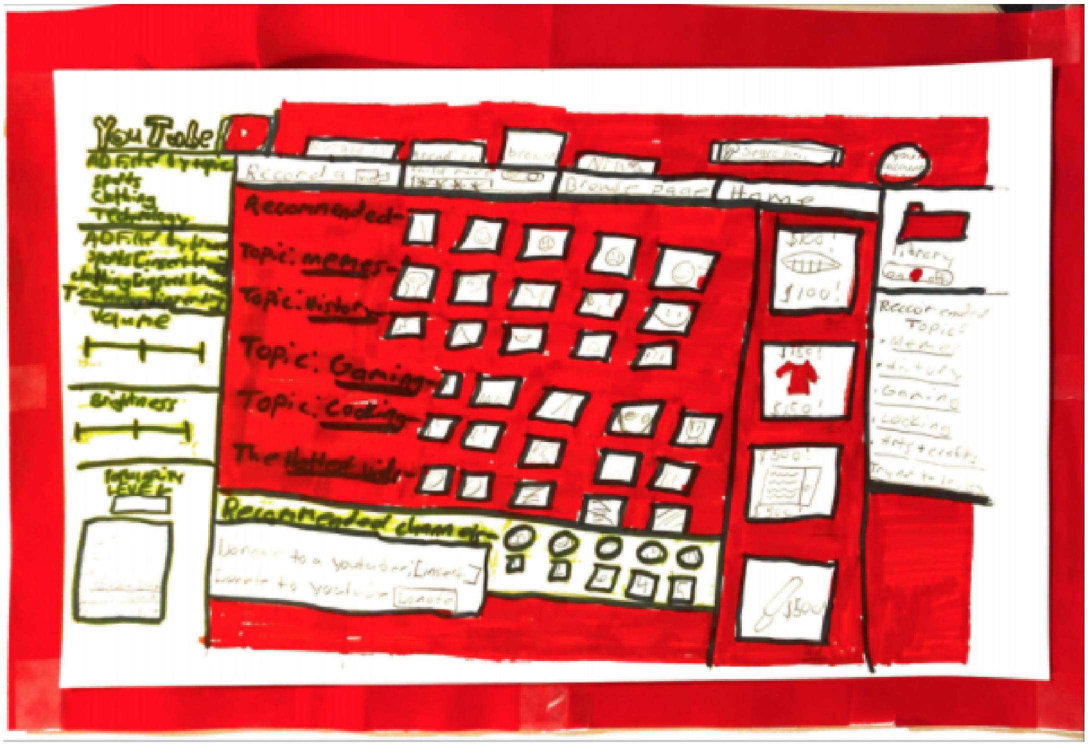


Figure 10-10: A paper prototype of a YouTube optimizing for entertainment and profit.

related to ads. Six out of the eleven projects added a child lock feature to their prototype to keep children away from unsafe content. Similarly, four pairs designed a feature to make video content safer (e.g., no violence or swear words). Four projects contained content filters so that the user could have more control over the content they were recommended. Three pairs had a feature such as a survey or comment box so that YouTube could use direct feedback from users to make design decisions instead of having an algorithm predict based on the users' passive behaviors. It is worth mentioning that some projects had several features in each category, especially with regards to displaying safe or appropriate content, such as in Case Study 2.

Socratic Seminar Discussion

The Socratic seminar was also an opportunity for students to apply the tools that they learned through the redesign project to a real situation in the technology industry. By both identifying and empathizing with various stakeholders, students were able to think critically about what YouTube ought to do, such as keeping the autoplay feature or separating out content of its young viewers. One student chose the safety of children as a top priority: "Yes [I think that YouTube Kids should be a separate product] because then even if YouTube doesn't make as much money as they used to, it's still important that kids don't watch grown-up stuff."

Another student decided that YouTube's financial health was most important, "I think that it would be a bad thing [to move YouTube Kids to a separate product]. Because, YouTube would be losing a lot of money. And the way that they could fix it is add a different setting. Maybe find a better way to do the restrictive mode, or something like that. Or add another setting that would help restrict, like a child mode setting that will only have, that would send them to a different part which is all kids' videos."

These discussions helped them move from passive users to conscientious consumers of YouTube. They were able to conjecture what specific design decisions YouTube might make, and how those decisions might benefit stakeholders such as YouTube, children, and parents. This helped them develop their own opinions regarding the proposed changes to YouTube.

Another student echoed this opinion but was more concerned about the financial effects this design would have on YouTubers, "I think they kind of should be the same thing because if someone, if they did separate, and someone decided they wanted to post a video on YouTube Kids and they weren't getting as many views as they wanted because there's not enough users or as many users on YouTube Kids, they don't have the choice, it's not never going to be on YouTube."

Some students, however, disagreed. Despite what many students feared about a potential decrease in quality and quantity of content, some believed that the content should be separated in order to protect children: "Because then even if YouTube doesn't make as much money as they used to, it's still important that kids don't watch grown-up stuff."

Regardless of their stance, students consistently referred to various stakeholder-value pairs when justifying their opinions.

10.3 What is the Effect of Learning about Ethics and AI?

Last, we examine what the effect of learning about artificial intelligence and ethics has on students. We explore how this curriculum may have changed their perception of AI as a technology and then how this curriculum may have changed their perception of themselves as empowered designers.

10.3.1 Perception of Artificial Intelligence

At the end of the week, students were once again asked to rate how strongly they agreed with several descriptions of AI: good, bad, complicated, simple, easy to understand, magical, dependable, factual, trustworthy, helpful, fair, biased, unbiased, objective, and subjective. A Wilcoxon Signed Rank Test was used to test for any significant shifts in students' agreement. Only two significant shifts were found: at

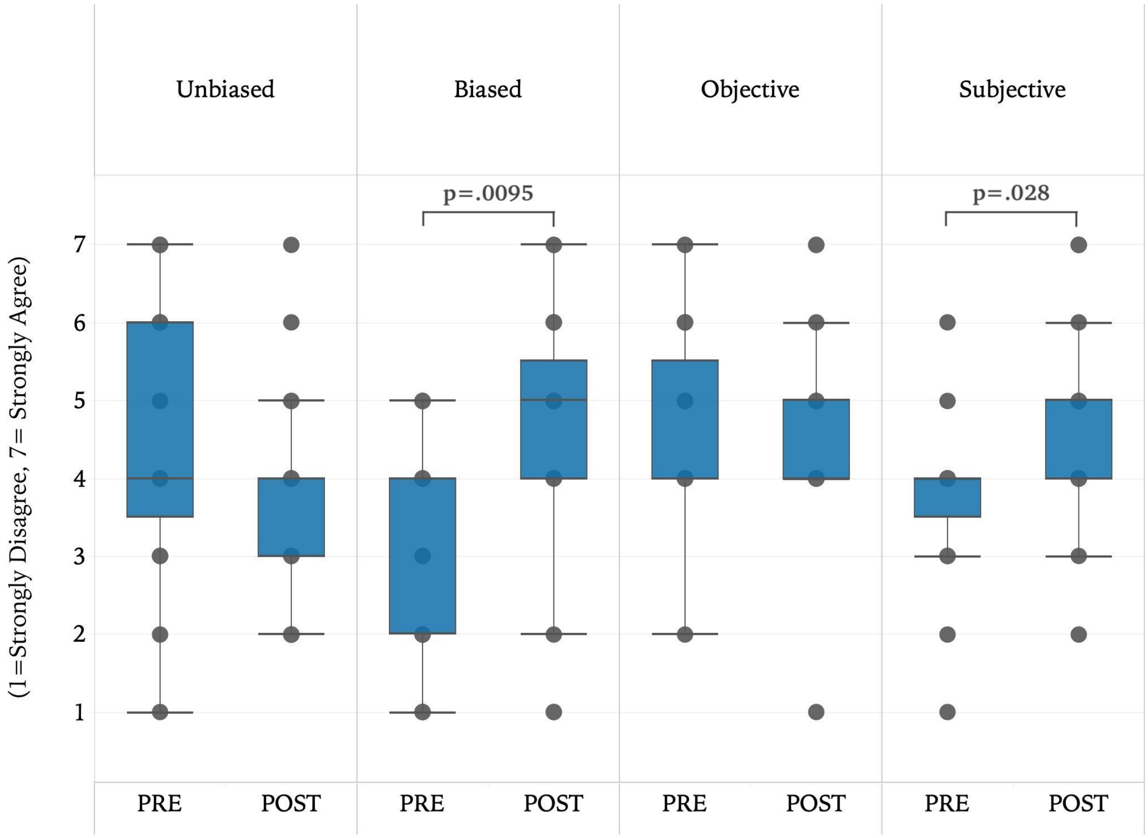


Figure 10-11: Shift in students' perceptions of AI as objective, subjective, unbiased, and biased by the end of the workshop.

the end of the week, students agreed more with the descriptions that AI is biased (with p=.0095) and that AI is subjective (p=.028). No significant shift in agreement or disagreement occurred with the terms “unbiased” or “objective.”

Additionally, at the end of the week, we asked students again about narratives around AI (such as immortality, ease, gratification, dominance, inhumanity, obsolescence, alienation, and uprising). At the end of the week, students showed no significant changes in their concern or excitement around each narrative nor were there any significant shifts in their thinking that these narratives were more likely to come true in their lifetimes.

10.3.2 Perceptions of Themselves as Empowered Designers

Finally, we investigate students’ perceptions of themselves in relation to artificial intelligence. Like in the beginning of the survey, in the final survey, we asked children how much they agreed with the statement, "I feel I am able to influence how Artificial Intelligence (AI) develops in the future." There was no significant shift in the response to this question.

Qualitatively, many students did voice an interest in being involved in future design processes and learning more about technology in general. Many students expressed interest in presenting their ideas regarding their YouTube redesign to various stakeholders. In a reflection after the activity, one student wrote, “I hope that our YouTube designs are actually considered in YouTube’s future plans.” Another student reflected a desire for their voice to be heard in the design process: “I hope that I get to work to help with the redesign on YouTube at Google.”

After the exercise, students also showed interest in learning more technical concepts to complement their design skills. Students originally came in with a wide range of exposure to technical concepts as well as practical experience with coding, and to some, the workshop was intimidating. However, at the end of the workshop, several students asked if there were other similar workshops they could enroll in with programming components. In their reflection, one student wrote, “How I’m going to make the AI?”, while another wrote, “I feel satisfied, yet hungry for more.”

"Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations."

Kate Crawford

11

Discussion

11.1 What Students Learned

Overall, there are several promising results from this study. First, students showed an upwards (though not statistically significant) trend in their ability to recognize artificial intelligence systems in their everyday life. For seven out of the ten technologies students were tested on, a greater percentage of students in the post-assessment correctly identified the technologies as AI or not. There were three technologies, however, where this was not the case.

The number of students who correctly identified a Nintendo Switch, a handheld gaming console, as not a form of artificial intelligence stayed the same from the beginning to the end of the week. This question was originally intended to test students' abilities to differentiate the concept of hardware from artificial intelligence.

However, there may have been some confusion from students about which aspect of the Nintendo Switch to which the question was referring. The Switch itself is not a form of AI; it is hardware. However, many of the apps that come standard with the Switch (such as the games store) do utilize AI and recommend games to users.

The other two technologies where students did worse at labeling them as artificial intelligence (or in this case, not as AI) were YouTube Subscriptions and Alexa Reminders. YouTube Subscriptions is a feature which automatically adds videos from particular YouTube channels a user selects to a list, while Alexa Reminders is an alarm application powered by Amazon. Neither service requires artificial intelligence. Both of these questions were included to get students to consider the definition of artificial intelligence given to them, in hopes that they would realize that neither a subscription nor reminder service would require a dataset to learn from or even a prediction to be useful. I attribute the decline in correct identification at the end of the week to two reasons. First, it is possible that students associate so strongly the brands of YouTube and Alexa with artificial intelligence that they identified these services as such. Second, especially in the case of Alexa Reminders, AI services surround the primary, non-AI feature. For example, to use Alexa Reminders, one must voice activate Alexa, which does use AI, even if setting an alarm does not.

In addition to an upward trend in their ability to recognize AI systems in the wild, this work also showed that middle school students have the capability to understand how supervised machine learning works. Interestingly, even students who did not answer correctly the supervised machine learning understanding check did answer correctly the questions about how to mitigate algorithmic bias. My instinct is to attribute this to students' exposure to computational thinking skills. The question which checked students' understanding of how supervised machine learning systems work required students to think about how machines think - what data has the computer seen before - which is different from how humans think. The solutions to the algorithmic bias questions are a bit more intuitive, although the answer to the second question about curating a dataset has two very similar multiple choice options which required students to differentiate the effects of each dataset.

Furthermore, the results show that students were able to identify statistically signifcantly more stakeholders for Google Search and YouTube at the end of the workshop. It is interesting that this is the case for these two technologies - the most popular in the class - and not self-driving cars. It is also interesting that in the original pilot workshop in Pittsburgh, there was not a significant shift in the number of stakeholders students could identify after the intervention either, where the technology in question was also self-driving cars. This implies that technologies familiar to children might be more effective in getting them to consider stakeholders, values, and follow ethical design protocols.

We also see a statistically significant shift in two descriptions of AI at the end of the week: students were more likely to agree with the descriptors "biased" and "subjective." Although one might expect there to be a downward shift in the descriptors "unbiased" and "objective," that was not the case. This indicates that students are more open to the possibility that AI can be subjective or biased, but perhaps do not see it as a descriptor which is true all of the time or for every system. It is worth noting that during both the pre- and post-assessments, students asked to use a dictionary to look up the words "objective" and "subjective."

11.2 Design Recommendations

Many lessons were learned throughout the process of developing the first and second iterations of this curriculum. The most important of those lessons are shared below.

11.2.1 What Worked

Several aspects of this curriculum were successful and are recommended for future iterations of this curriculum or similar projects. These are discussed below.

Define and Clarify Terminology

One finding from the original pilot was that students were often unclear about terminology. The meanings and differences between terminology such as artificial intelligence, algorithm, robot, electronics, programming, code, etc., were unclear and often incorrectly exchanged. The very first activity in this curriculum (AI Bingo) takes the time to explain, in a very simple way, what artificial intelligence is and how it relates to the term "algorithm" as well as "robot." This lesson set up students for success later, as there is an upward trend by the end of the week in students' ability to identify everyday AI systems. Dedicating an activity solely to clarifying terminology has proven to be useful to educators already. This activity has been translated into Portuguese and was featured by MIT Technology Review as an activity for parents to do with their children.

Situate Students as Experts

Another design recommendation is to situate students as experts in the classroom. Eighteen out of 19 students reported regularly using YouTube and Google Search, both of which were technologies students were able to identify statistically significantly more stakeholders for in the post-assessment. During the Socratic seminar, students were engaged in the discussion because it was centered around YouTube, and students were also incredibly detailed when completing the YouTube Scavenger Hunt. For example, instead of simply listing the most commonly known AI systems associated with YouTube, students identified very niche AI systems, such as auto-captioning system or the copyright infringement detection algorithm. This recommendation is also consistent with the Gradual Release of Responsibility model mentioned in a previous chapter, which states that all lessons should begin with an explicit connection between what students will be learning and what they are already familiar with.

Provide "Real World" Motivation

Similarly, design recommendation is to provide students with "real world" motivations for what they are learning. At the beginning of the summer workshop, students seemed skeptical of what they were learning (perhaps in part to the expectations gap mentioned in the “What Didn't Work” section). However, the third day of camp proved to be a turning point. On Wednesday of that week, Joy Boulamwini (whose research the algorithmic bias activity is based upon and whom the students had watched in a video associated with that activity) testified before congress about algorithmic bias in facial recognition systems. This led students to write questions like, "Why does congress not know about today's modern tech?" or "I hope that facial recognization [sic] changes so everyone can be abled [sic] to get recognized." on the classroom mural. Students were also generally more engaged in classroom discussions from that point on.

Provide Actionable Ethical Design Protocols

One highly successful aspect of this curriculum is its usage of the ethical matrix as an ethical design tool. The ethical matrix provided a relatively quick and straightforward way to start conversations around the ethics of AI design. Anytime a new form of AI was being discussed, students immediately knew to think first of the possible stakeholders associated with the technology and second to think of the values that might be embedded in that AI system. While many students used the ethical matrix in a utilitarian way to decide what their YouTube redesigns should optimize for (e.g., choosing the value that was most popular among all stakeholders), several groups of students used the ethical matrix as an actionable exercise which led to a more abstract discussion of values. Several groups even modified their ethical matrix, using different colors or patterns within the matrix to indicate varying levels of conflicts or importance in values.

Future AI and ethics curricula should include easy to follow, actionable ethical design protocols or exercises like the ethical matrix.

11.2.2 What Didn't Work

Not every aspect of this curriculum's design was successful. In this section, the challenges of those aspects are identified.

Format Is Everything

One major tension in this work is what format is best suited for this kind of curriculum. Through both iterations, the pros and cons of teaching in a traditional classroom setting versus a summer workshop became very evident. There are two main benefits of teaching in a traditional classroom setting: students are more engaged at the risk of getting a bad grade, and there is very little attrition (which, depending on the location of the school, could also mean a greater diversity of students will have access to the curriculum). There are also two major disadvantages to the traditional classroom setting: the challenge of bringing politics into the classroom (due to the fact that one of the primary objectives of this curriculum is to show how seemingly neutral artifacts are inherently political) and the challenge of finding the time to cover non-standardized or required topics.

A summer workshop format, like the second iteration of this curriculum, solves both of the problems with the traditional classroom setting but runs into the issue where there is more likely to be higher attrition (for example, several students left the workshop two days early for their family summer vacation) and a more homogenous group.

Neither format is perfect, and future workshops should carefully consider how to mitigate these problems. An after school program might be a compromise format between the two.

Student Expectations of AI

The decision to use unplugged activities was an early design decision based on interviews with educators who were intimidated by the idea of teaching AI and based on the idea that there should not be any costly technological barrier to prevent vulnerable communities from learning about the ethics of AI. There are, however, a few downsides to utilizing unplugged activities.

Many students attended the workshop expecting to work with robots or some other form of cutting-edge technology. This means that for the first two days of camp, students were not as engaged or enthusiastic as they might have otherwise been. Additionally, some students were disappointed to find out that the workshop would not focus on how to prevent the "robot uprising." This mismatch in student expectations and the focus of the curriculum led to some of the attrition throughout the week.

Future iterations of this curriculum might involve more forms of technology to meet the expectation of students. Given the fact that there was no significant shift regarding common narratives around AI from the beginning of the week to the end of the week, it would probably be worthwhile to have a specific activity address or debunk those common AI tropes.

11.3 Study Limitations

This work is limited in several ways, which provide an opportunity for further investigation. The first limitation of this work is the small sample size and the setting in which the study took place. The study took place during summer break, and so the attitudes towards class were different than in a typical classroom. Given that students were enjoying their summer vacation, many were in and out of the classroom during the week, which made continuity difficult and resulted in a decrease in data points. Additionally, since the pre-assessment was administered before any workshop activities, and the post-assessment was given at the very end, it is hard to pinpoint the effects of individual activities on students. Similarly, given that it was a summer camp, other activities such as campus or lab tours might have influenced students' understanding and perception.

A second limitation is that students were also broken up into two classrooms by age, which meant two separate experiences for the older and younger students. While instructors for both classes followed the same script, classroom management style varied out of necessity. Originally, this split by age was by design in order to better understand how different age groups would engage with the curriculum, but given the issue of attrition due to the summer camp format, the sample sizes were very small.

A third limitation is that the participants were a very homogeneous group which self-selected into the program. Unfortunately, there was a large gender imbalance due to the fact that recruiting was handled through a STEM programming company, which often skews male. This is deeply regrettable given the empahsis in the curricula on the importance of diversity. The students were also predominantly of high socio-economic status, as evidenced by the fact that the camp took place on MIT's campus and required child pickup at 4:00 p.m. each afternoon.

11.4 Future Work

In addition to overcoming the limitations of this study, there is an incredible amount of future work to be done. Several questions remain, such as:

1. What kind of professional development can best support educators who wish to use this curriculum?
2. Which age groups are best suited for each activities?
3. What format is the best way to deliver this curriculum - in school, during after school programs, summer camp, or something else?

Many additions to the content could also be made. For example, future work might include additional kinds of artificial intelligence systems such as rules-based systems or generative adversarial networks and exploring the ethical ramifications associated with those algorithms. Other meaningful additions might explore how workplace dynamics affect the "opinions" algorithms hold or tie in with civics lessons and theory of social change. Education about non-traditional AI careers, such as policymaking, community organizing, or the social sciences, could also potentially assist students in finding greater agency around creating ethical AI. The best change in content one could bring to this curriculum is finding a way to empower students to conduct participatory design around AI with a diverse set of stakeholders.

Due to the homogenous nature of the participants in this study, it is important that future work be broadened to a more diverse classroom.

"Just because you find that life's not fair, it doesn't mean that you just have to grin and bear it.... Even if you're little you can do a lot, you mustn't let a little thing like little stop you...."

"Naughty" from *Matilda the Musical* based on the book by Roald Dahl

12 Conclusion

12.1 Answers to Research Questions

In this section, we revisit the research questions posed at the beginning of this thesis.

Which concepts in the realm of AI and ethics should be prioritized for such a curriculum?

In order for children to become conscientious consumers, ethical designers, and democratic participants around AI, three objectives were prioritized in this curriculum:

1. Getting students to recognize AI in their everyday life.
2. Showing students that AI is neither an objective source of information nor morally neutral.

3. Giving students ethical design tools which they can apply to any socio-technical system, including AI.

Both concepts, 1 and 2, are motivated by the need for the children of today, who are "AI natives," to be conscientious consumers of artificial intelligence. While the need for the public at large to understand that artificial intelligence, or technology in general, is not morally neutral is underscored by a rich literature, the initial pilot of this curriculum showed that this concept is only useful if children can recognize AI in their everyday life.

The third concept is prioritized for a number of reasons. One of these reasons is that we do not want the children of today, the professional technologists of tomorrow, to repeat the mistakes of the past. However, beyond this reason, knowledge of ethical design tools will empower children as citizens in the era of AI. Beyond AI literacy, we must give students tools to be fluent in the ethical design of artificial intelligence.

What are children's preconceptions of artificial intelligence and the ethical issues associated with it?

This work showed that students' preconceptions of AI were largely based around embodied or human-like agents. Before the intervention, most students associated AI with robots even though their most commonly used forms of AI are entirely software-based, such as Google Search or YouTube. Many students were able to acknowledge that AI was programmed or made by a human, but most could not describe technically how an AI system might work or the role of data.

This work also showed that students are familiar with many popular narratives around the societal impact of artificial intelligence, especially narratives around the automation of jobs or the idea that AI will make everyday tasks easier for humans. Despite their familiarity with these narratives, students were initially unable to, on average, name more than a couple of stakeholders associated with AI.

What are children capable of learning about key concepts in AI and ethics? How does this inform a developmentally appropriate curriculum?

Children were able to learn several important concepts about artificial intelligence and ethics. This work showed that students were able to recognize various forms of AI in their everyday life. Beyond that, students were capable of understanding the basic mechanics of supervised machine learning and could understand the role a dataset plays in the outcome of such a system.

This work also showed that students were capable of considering the effects of AI designs on various stakeholders, and could use those considerations to re-imagine popular technologies.

How can a curriculum be designed so that it is useful, accessible, supports educators, and is effective in student learning?

The original design, pilot, and subsequent iteration of this curriculum were very informative regarding how future curricula can be designed to address the topic of AI and ethics, and ethics and technology in general. When designing future curricula, this work suggests that best practices include taking time to clarify terminology, situating students as experts, providing students with "real world" motivation for what they are learning, and providing actionable ethical design exercises for students to practice.

How does learning about the ethical design of artificial intelligence change children's perceptions of artificial intelligence?

In this work, we saw students shift from seeing artificial intelligence only as objective and unbiased to being both objective and subjective as well as both unbiased and biased.

How does learning about the ethical design of artificial intelligence change their perceptions of themselves as empowered designers?

While the results from the second workshop did not indicate any shift in students' perception of their influence over AI design, students in both workshops showed an interest in furthering their education in AI and in ethical design. Furthermore, students in both workshops indicated that they wanted their voices to be heard by big technology companies.

12.2 Final Thoughts

At the beginning of this thesis, I asked the question: how do we ensure that the children of today have the same advantages regarding technology that my friend Maribeth had when searching for a job? If artificial intelligence is going to be pervasive across society (and arguably, it already is), it is imperative that we democratize power over that technology, including and especially for society's youngest members.

This work shows that the "AI natives" of today are not only capable of understanding how these technical systems work but also capable of participating in the ethical design of AI systems.

During the first day of the pilot workshop, as students were working on the first draft of their algorithm for the "best" peanut butter and jelly sandwich, a student raised his hand and patiently waited for me to come help him. I walked over and leaned in to hear his question.

"I'm wondering," he asked, "can my algorithm be my opinion?" I smiled.

That's a start.

A
Curriculum

The open-source curriculum developed for this thesis can be found at http://bit.ly/mit-ai-ethics under the CC-BY-NC license.

The linked document contains instructor guides for all eight activities, the length of time each activity takes, any relevant worksheets or print-outs, and related slides. Creative commons guidelines are also included for those wishing to tweak, remix, or build upon this work.

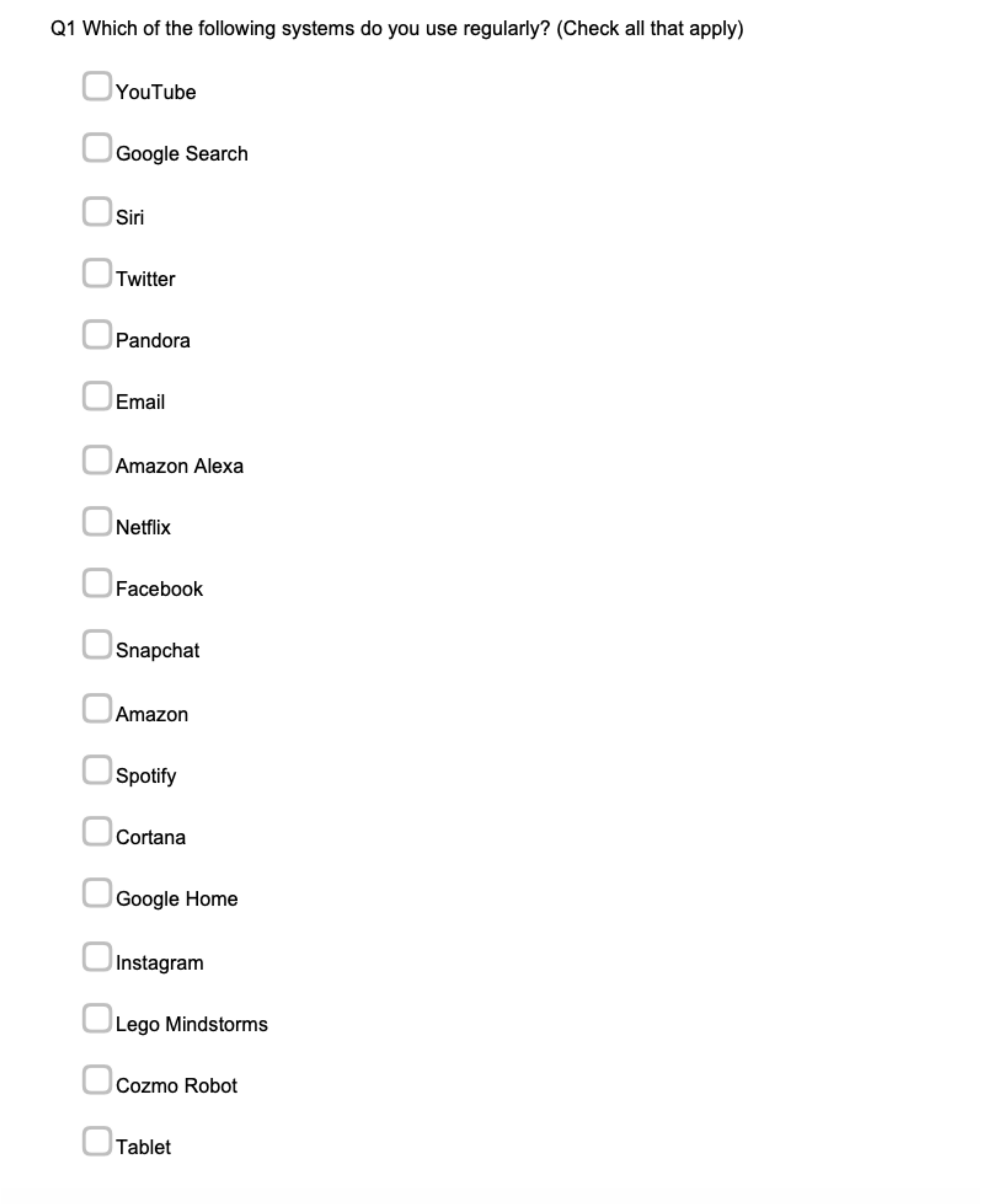
Additionally, the curriculum has been translated into German, Portuguese, and Korean and those translations are also linked in the document.

B
Assessments

The following assessments were used throughout the summer workshop pilot.

B.1 Pre- and Post-Assessment Survey

The following survey, in part based on the survey presented in [40], was used as both a pre-workshop and post-workshop survey. Each block of questions related to narratives surrounding AI were randomized in the order that they appeared to each participant.



Q1 Which of the following systems do you use regularly? (Check all that apply)

□ YouTube

□ Google Search

□ Siri

□ Twitter

□ Pandora

□ Email

□ Amazon Alexa

□ Netflix

□ Facebook

□ Snapchat

□ Amazon

□ Spotify

□ Cortana

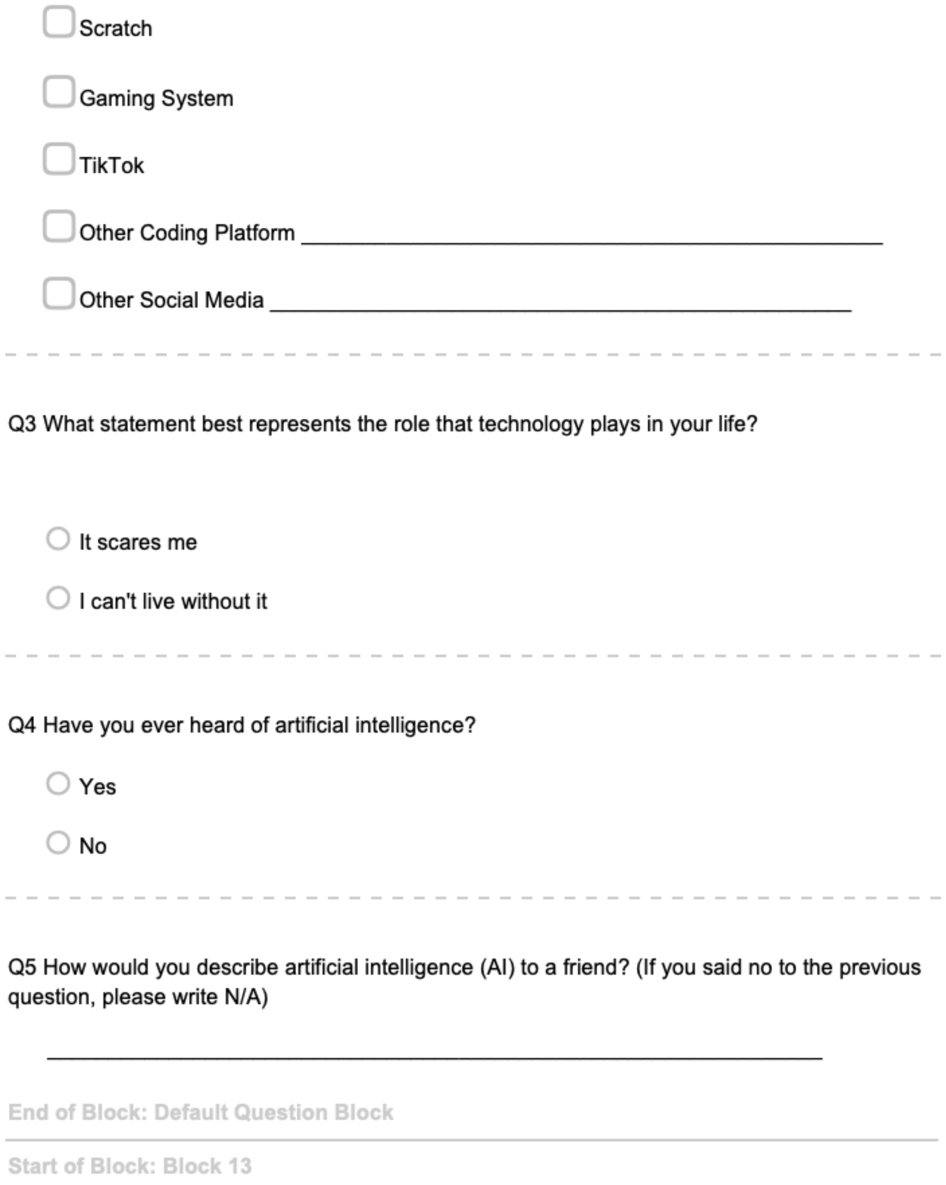
□ Google Home

□ Instagram

□ Lego Mindstorms

□ Cozmo Robot

□ Tablet



□ Scratch

□ Gaming System

□ TikTok

□ Other Coding Platform ____

□ Other Social Media ____

Q3 What statement best represents the role that technology plays in your life?

○ It scares me

○ I can't live without it

Q4 Have you ever heard of artificial intelligence?

○ Yes

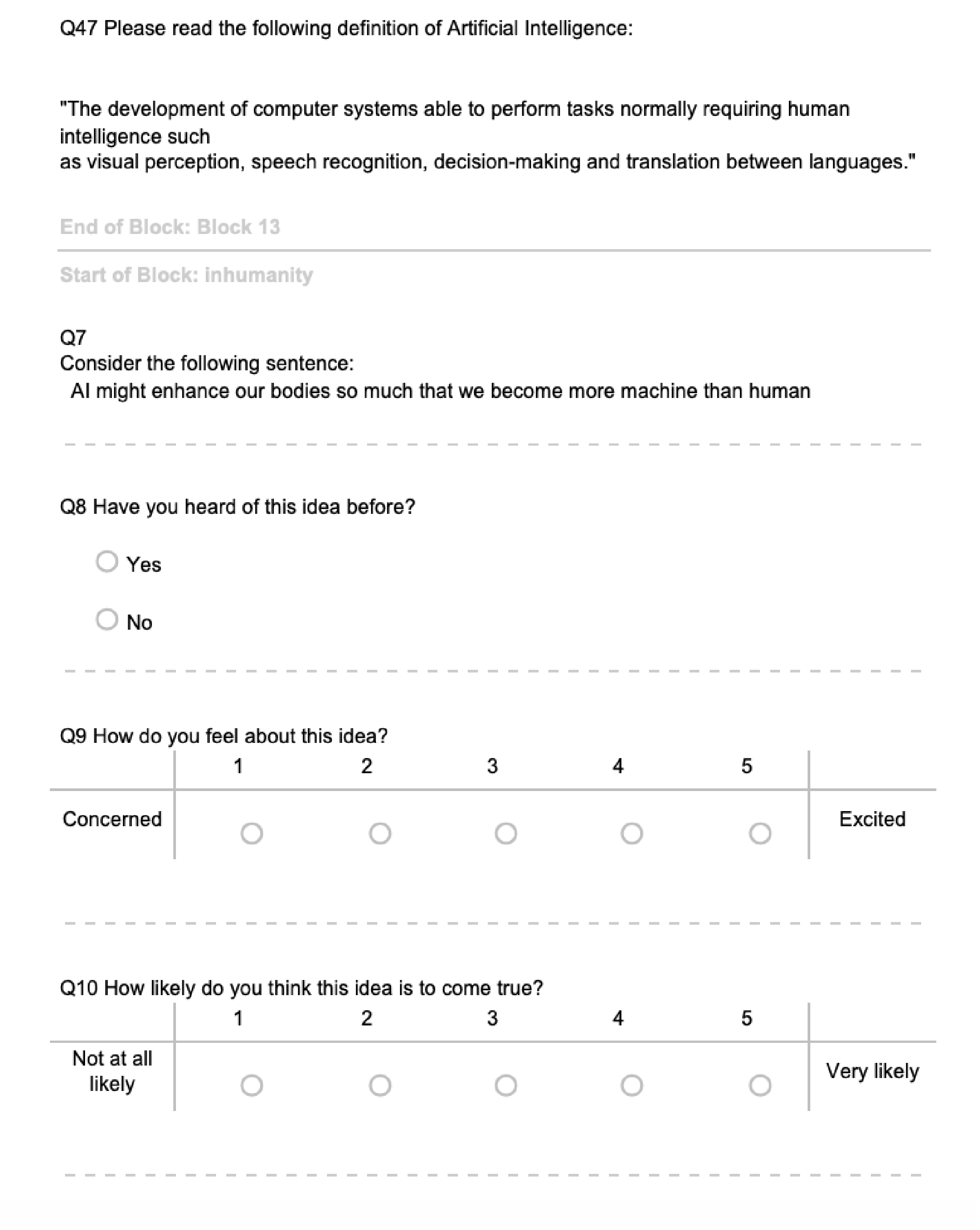
○ No

Q5 How would you describe artificial intelligence (AI) to a friend? (If you said no to the previous question, please write N/A)

____

End of Block: Default Question Block

Start of Block: Block 13



Q47 Please read the following definition of Artificial Intelligence:

"The development of computer systems able to perform tasks normally requiring human intelligence such
as visual perception, speech recognition, decision-making and translation between languages."

End of Block: Block 13

Start of Block: inhumanity

Q7
Consider the following sentence:
AI might enhance our bodies so much that we become more machine than human

Q8 Have you heard of this idea before?

○ Yes

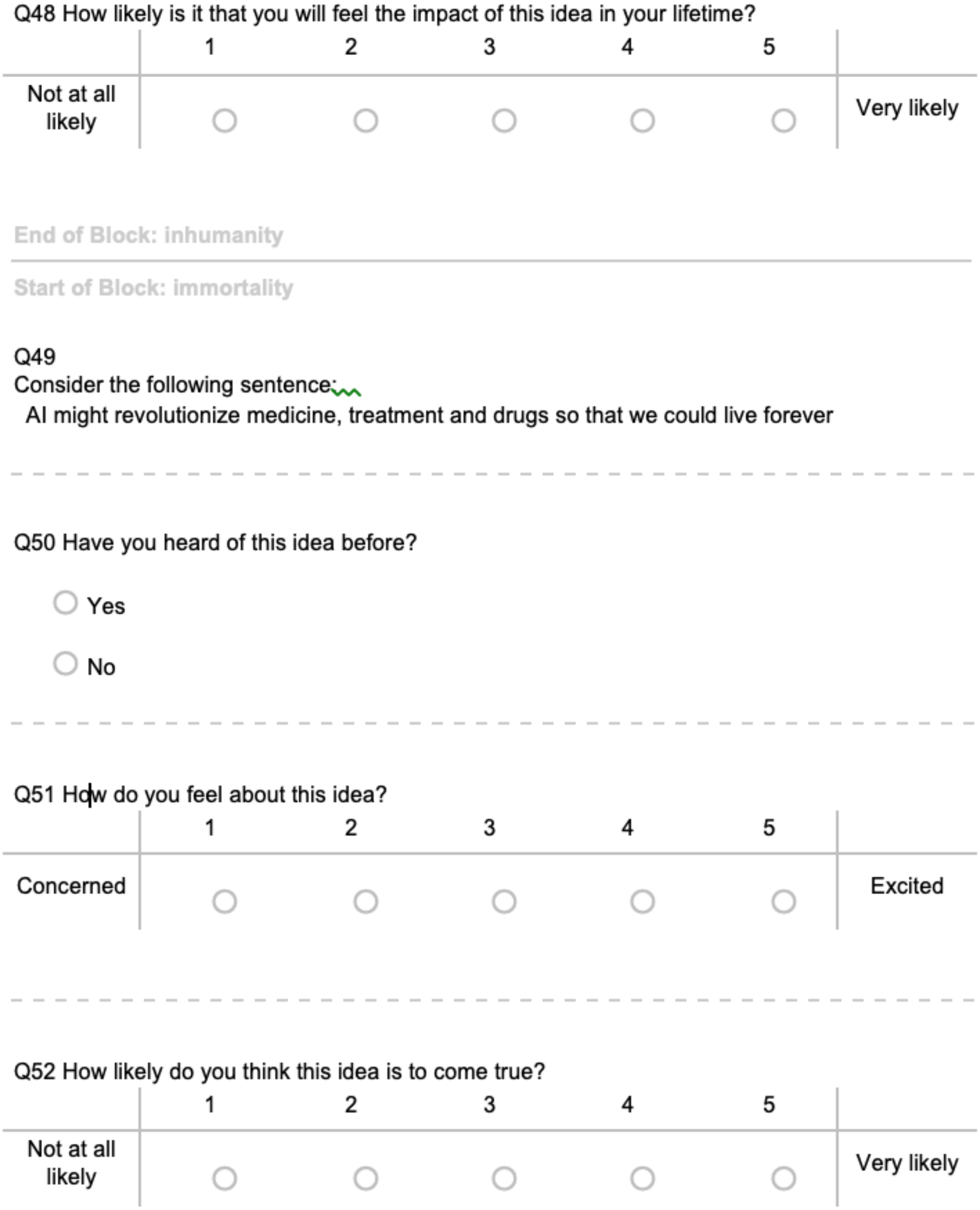
○ No

Q9 How do you feel about this idea?

	1	2	3	4	5	
Concerned	○	○	○	○	○	Excited

Q10 How likely do you think this idea is to come true?

	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely



Q48 How likely is it that you will feel the impact of this idea in your lifetime?

	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely

End of Block: inhumanity

Start of Block: immortality

Q49
Consider the following sentence:
AI might revolutionize medicine, treatment and drugs so that we could live forever

Q50 Have you heard of this idea before?

○ Yes

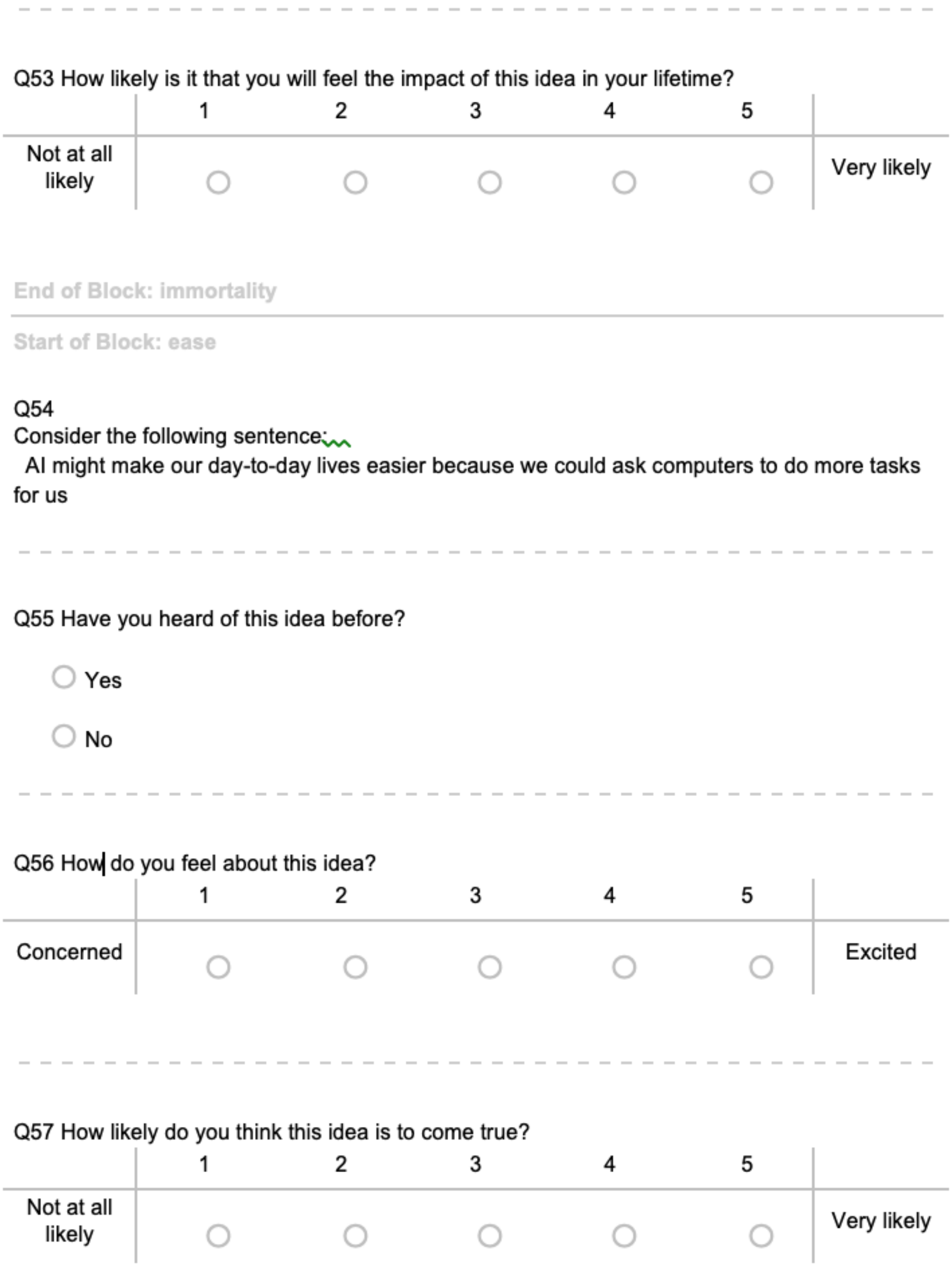
○ No

Q51 How do you feel about this idea?

	1	2	3	4	5	
Concerned	○	○	○	○	○	Excited

Q52 How likely do you think this idea is to come true?

	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely



Q53 How likely is it that you will feel the impact of this idea in your lifetime?

	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely

End of Block: immortality

Start of Block: ease

Q54
Consider the following sentence:
AI might make our day-to-day lives easier because we could ask computers to do more tasks for us

Q55 Have you heard of this idea before?

○ Yes

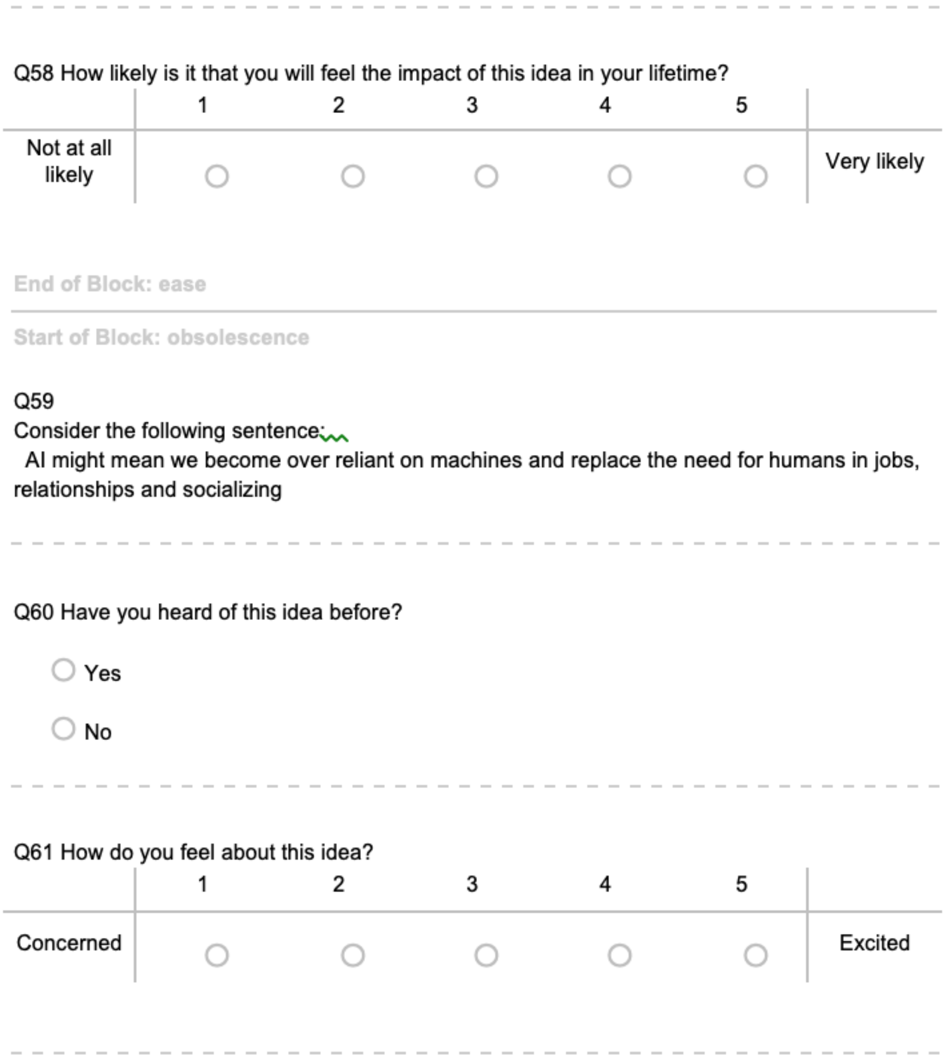
○ No

Q56 How do you feel about this idea?

	1	2	3	4	5	
Concerned	○	○	○	○	○	Excited

Q57 How likely do you think this idea is to come true?

	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely



Q58 How likely is it that you will feel the impact of this idea in your lifetime?

	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely

End of Block: ease

Start of Block: obsolescence

Q59
Consider the following sentence:
AI might mean we become over reliant on machines and replace the need for humans in jobs, relationships and socializing

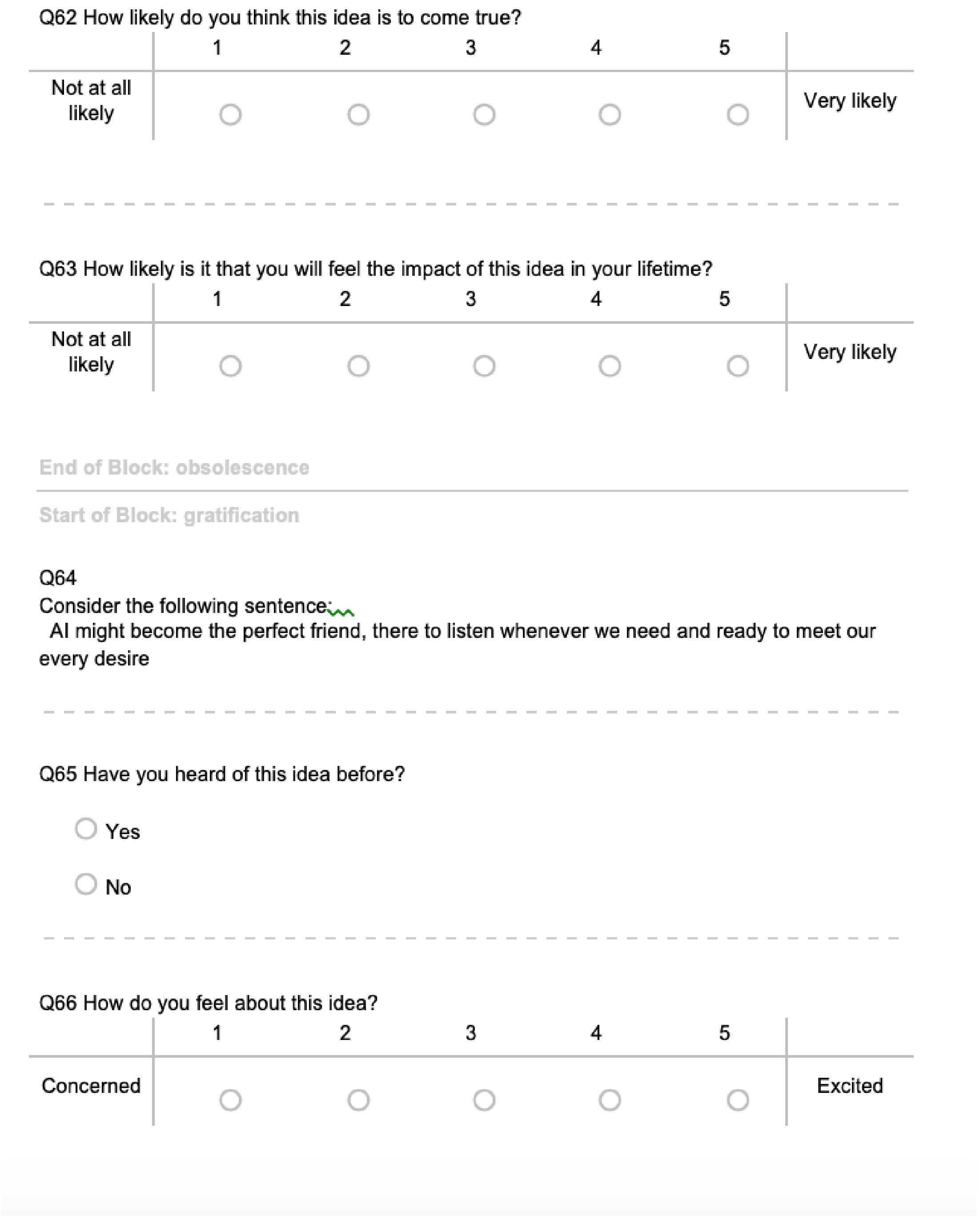
Q60 Have you heard of this idea before?

○ Yes

○ No

Q61 How do you feel about this idea?

	1	2	3	4	5	
Concerned	○	○	○	○	○	Excited



Q62 How likely do you think this idea is to come true?

	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely

Q63 How likely is it that you will feel the impact of this idea in your lifetime?

	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely

End of Block: obsolescence

Start of Block: gratification

Q64
Consider the following sentence:
AI might become the perfect friend, there to listen whenever we need and ready to meet our every desire

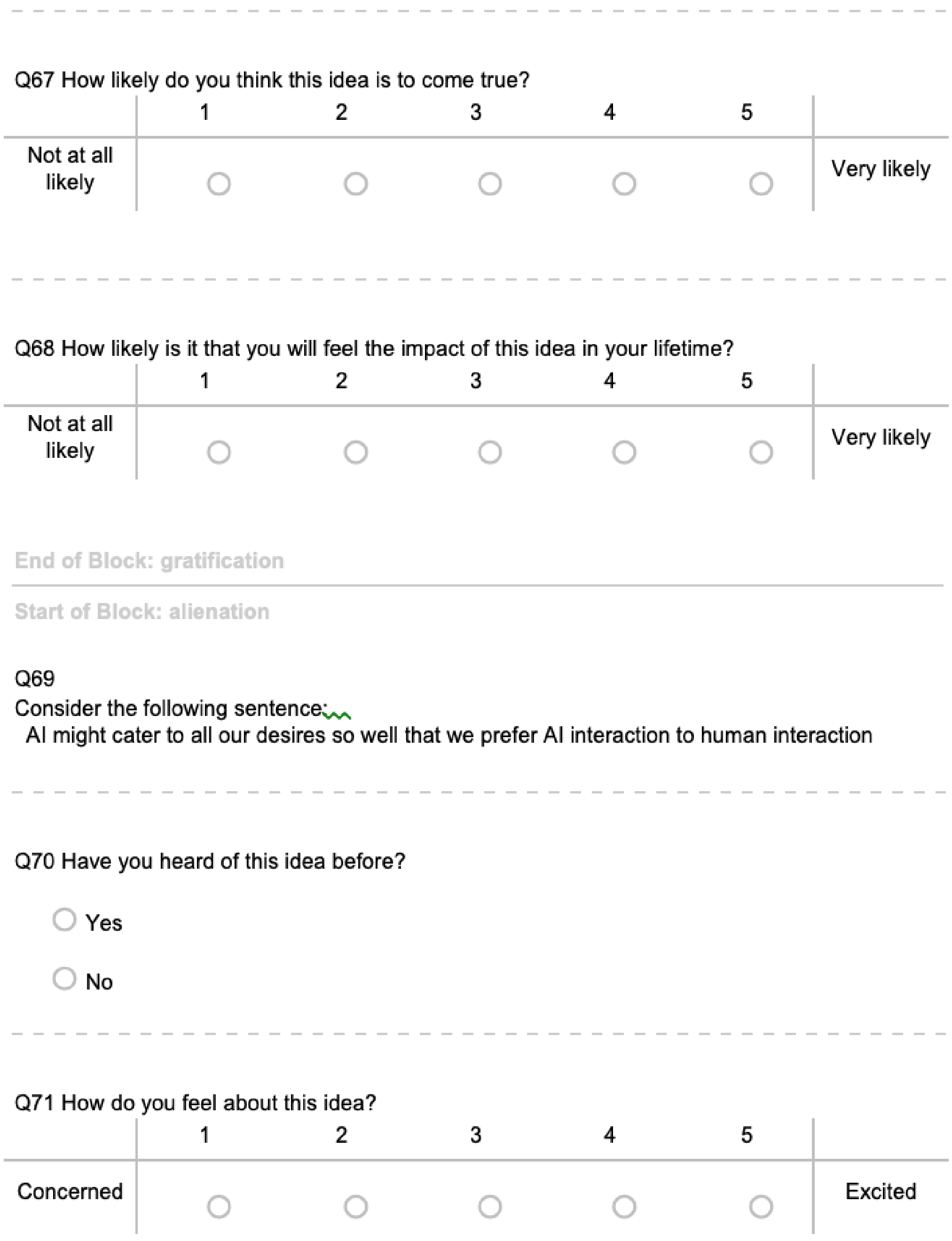
Q65 Have you heard of this idea before?

○ Yes

○ No

Q66 How do you feel about this idea?

	1	2	3	4	5	
Concerned	○	○	○	○	○	Excited



Q67 How likely do you think this idea is to come true?

	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely

Q68 How likely is it that you will feel the impact of this idea in your lifetime?

	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely

End of Block: gratification

Start of Block: alienation

Q69
Consider the following sentence:
AI might cater to all our desires so well that we prefer AI interaction to human interaction

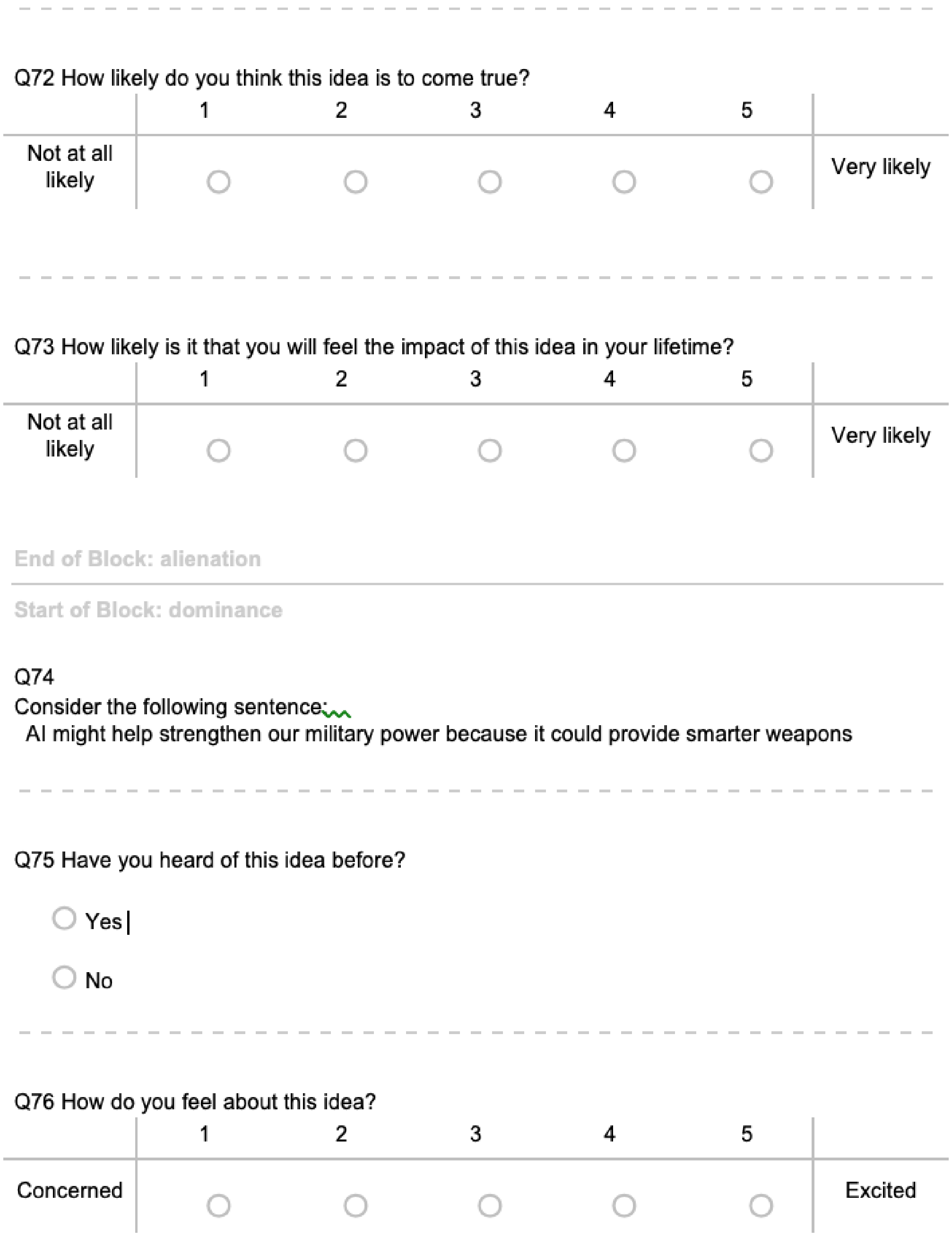
Q70 Have you heard of this idea before?

○ Yes

○ No

Q71 How do you feel about this idea?

	1	2	3	4	5	
Concerned	○	○	○	○	○	Excited



Q72 How likely do you think this idea is to come true?

	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely

Q73 How likely is it that you will feel the impact of this idea in your lifetime?

	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely

End of Block: alienation

Start of Block: dominance

Q74
Consider the following sentence:
AI might help strengthen our military power because it could provide smarter weapons

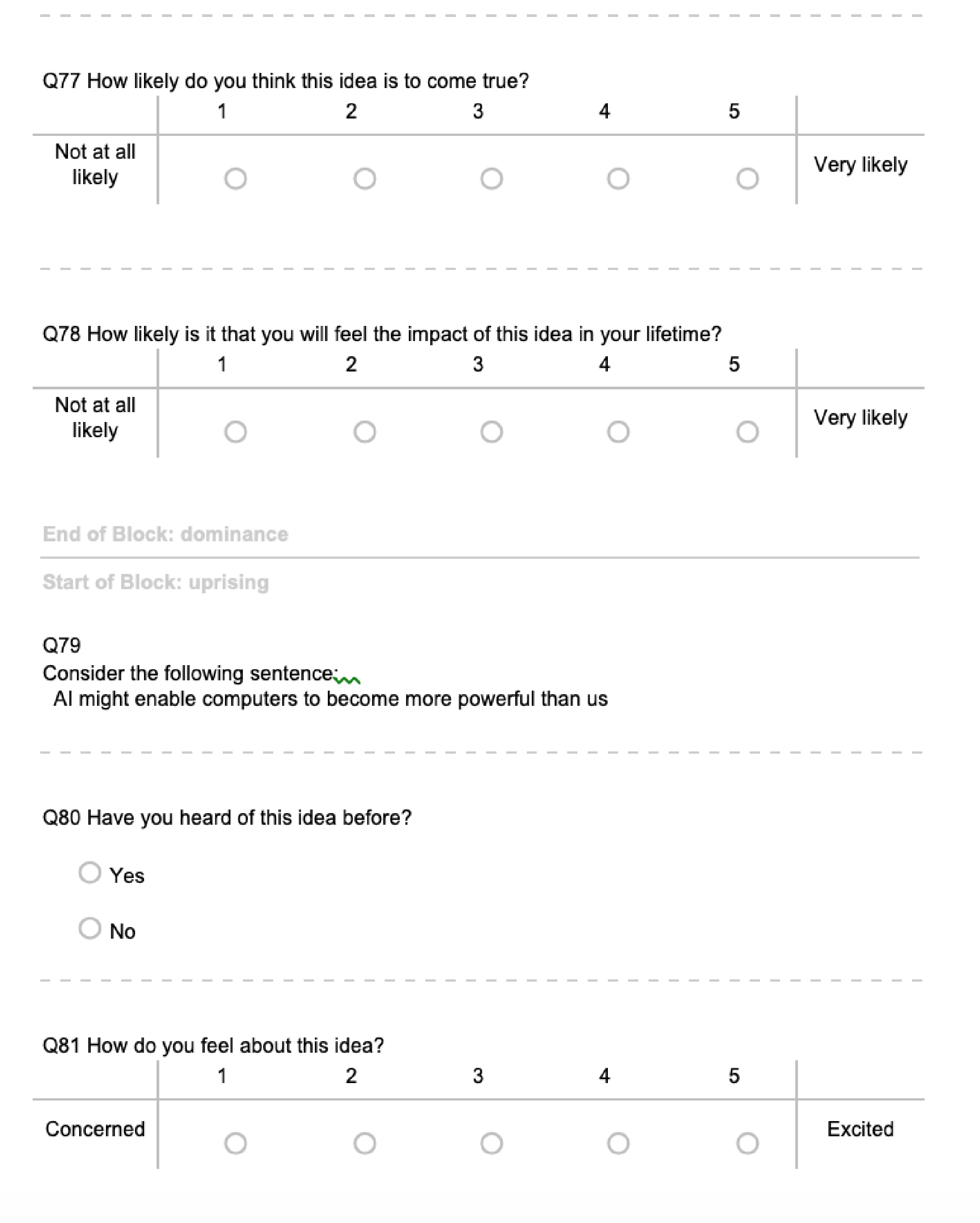
Q75 Have you heard of this idea before?

○ Yes

○ No

Q76 How do you feel about this idea?

	1	2	3	4	5	
Concerned	○	○	○	○	○	Excited



Q77 How likely do you think this idea is to come true?

	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely

Q78 How likely is it that you will feel the impact of this idea in your lifetime?

	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely

End of Block: dominance

Start of Block: uprising

Q79
Consider the following sentence:
AI might enable computers to become more powerful than us

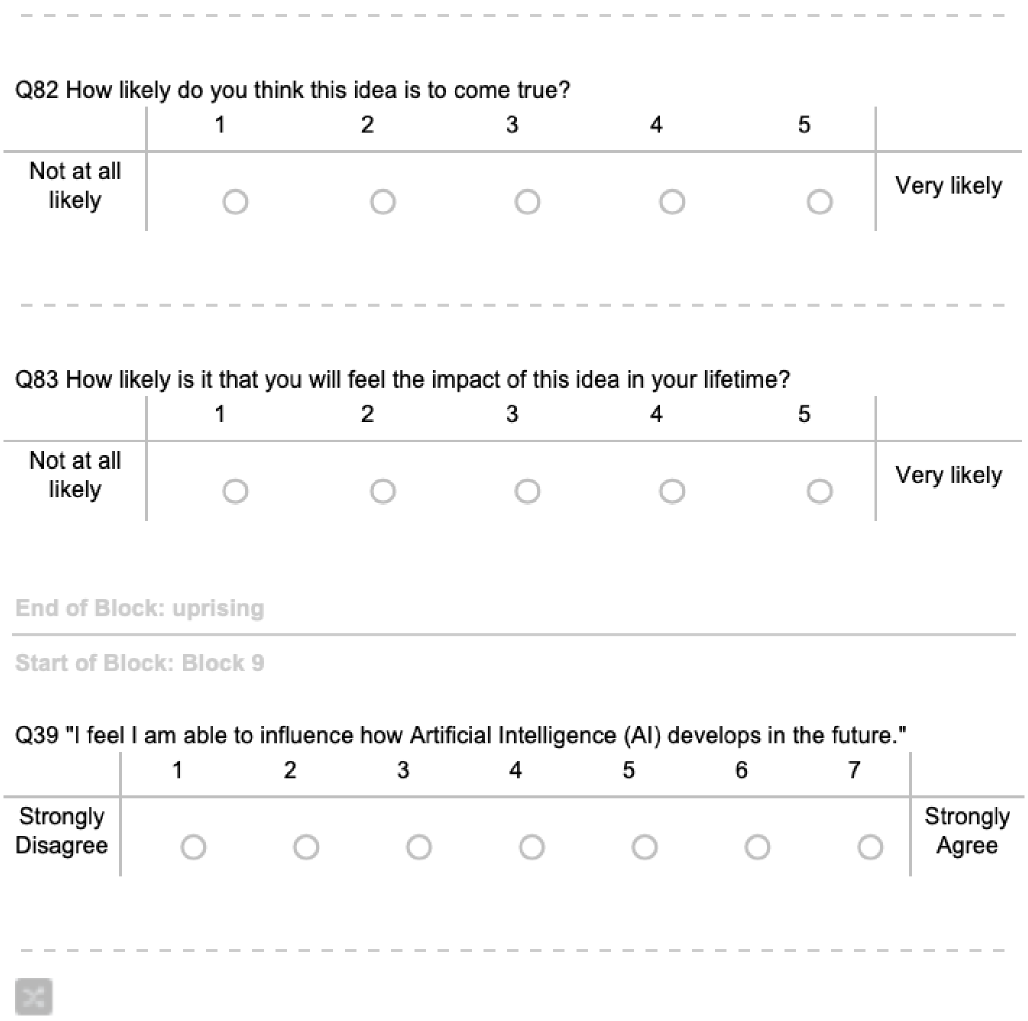
Q80 Have you heard of this idea before?

○ Yes

○ No

Q81 How do you feel about this idea?

	1	2	3	4	5	
Concerned	○	○	○	○	○	Excited



Q82 How likely do you think this idea is to come true?

	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely

Q83 How likely is it that you will feel the impact of this idea in your lifetime?

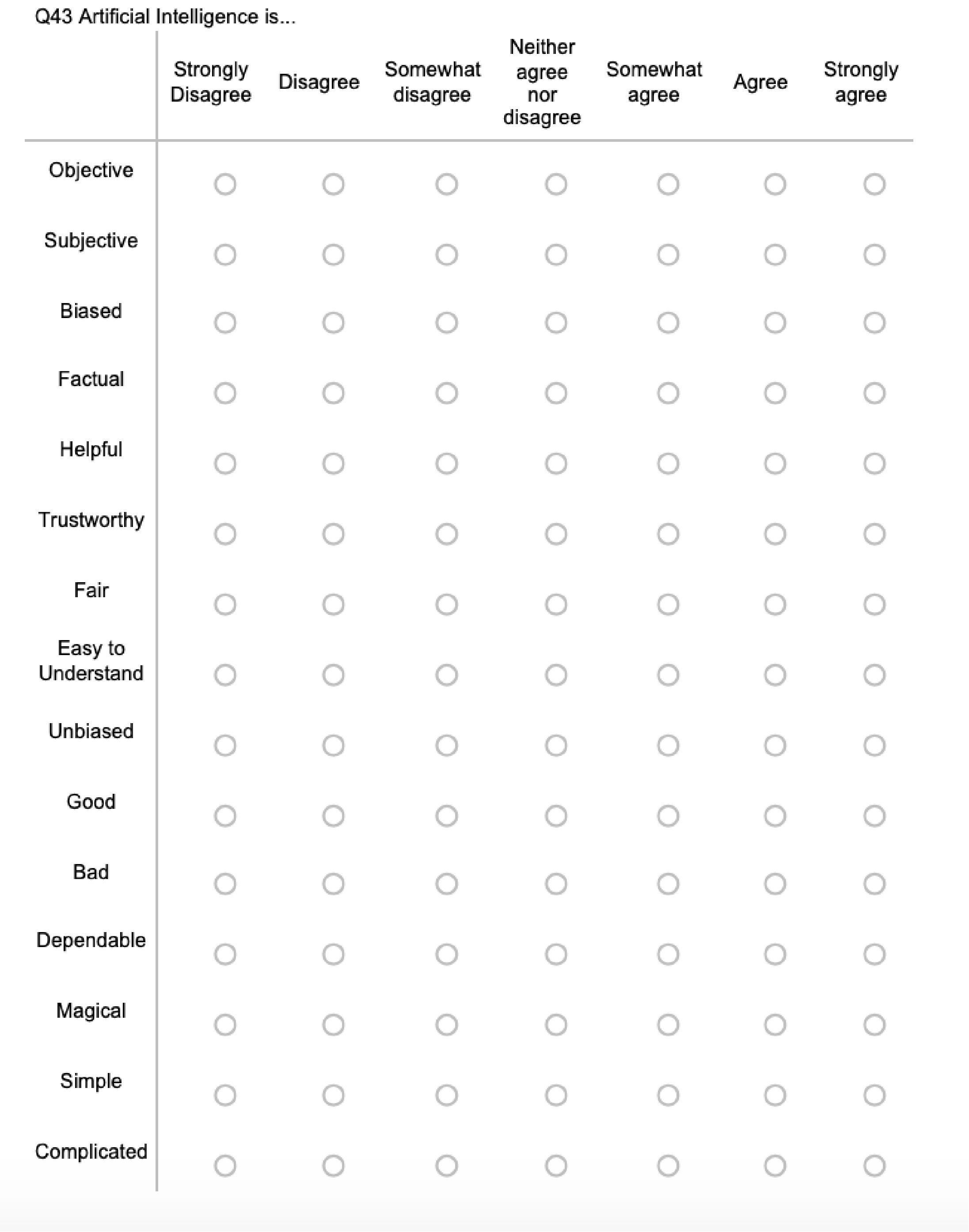
	1	2	3	4	5	
Not at all likely	○	○	○	○	○	Very likely

End of Block: uprising

Start of Block: Block 9

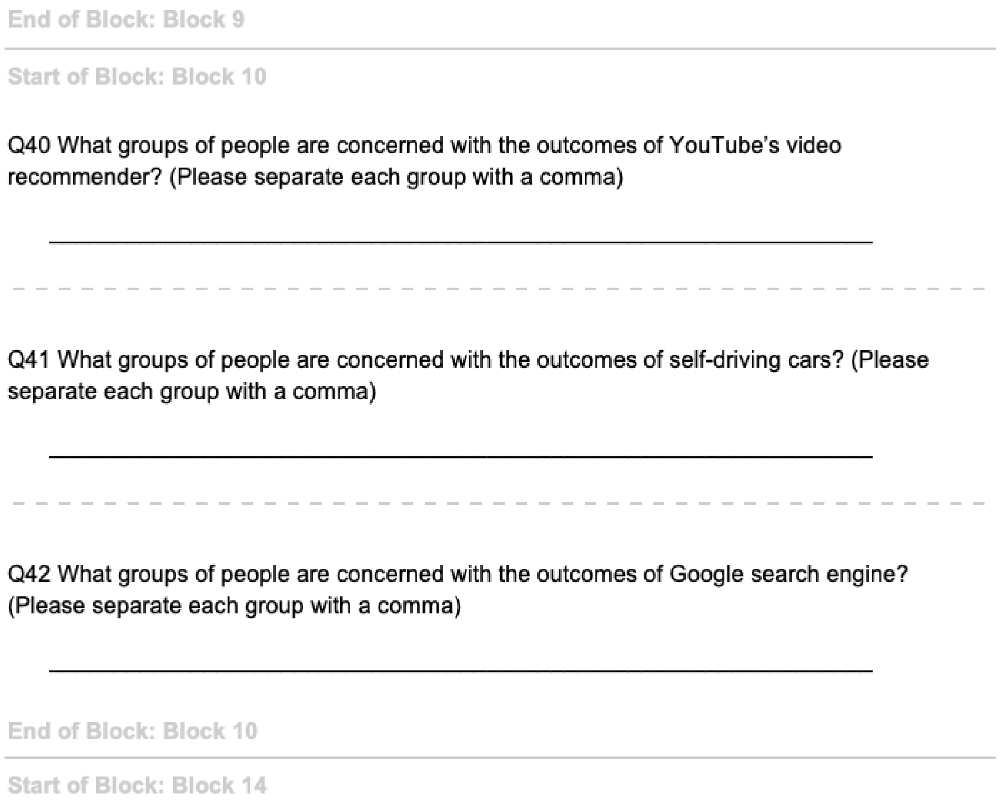
Q39 "I feel I am able to influence how Artificial Intelligence (AI) develops in the future."

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree



Q43 Artificial Intelligence is...

	Strongly Disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
Objective	○	○	○	○	○	○	○
Subjective	○	○	○	○	○	○	○
Biased	○	○	○	○	○	○	○
Factual	○	○	○	○	○	○	○
Helpful	○	○	○	○	○	○	○
Trustworthy	○	○	○	○	○	○	○
Fair	○	○	○	○	○	○	○
Easy to Understand	○	○	○	○	○	○	○
Unbiased	○	○	○	○	○	○	○
Good	○	○	○	○	○	○	○
Bad	○	○	○	○	○	○	○
Dependable	○	○	○	○	○	○	○
Magical	○	○	○	○	○	○	○
Simple	○	○	○	○	○	○	○
Complicated	○	○	○	○	○	○	○



End of Block: Block 9

Start of Block: Block 10

Q40 What groups of people are concerned with the outcomes of YouTube's video recommender? (Please separate each group with a comma)

____

Q41 What groups of people are concerned with the outcomes of self-driving cars? (Please separate each group with a comma)

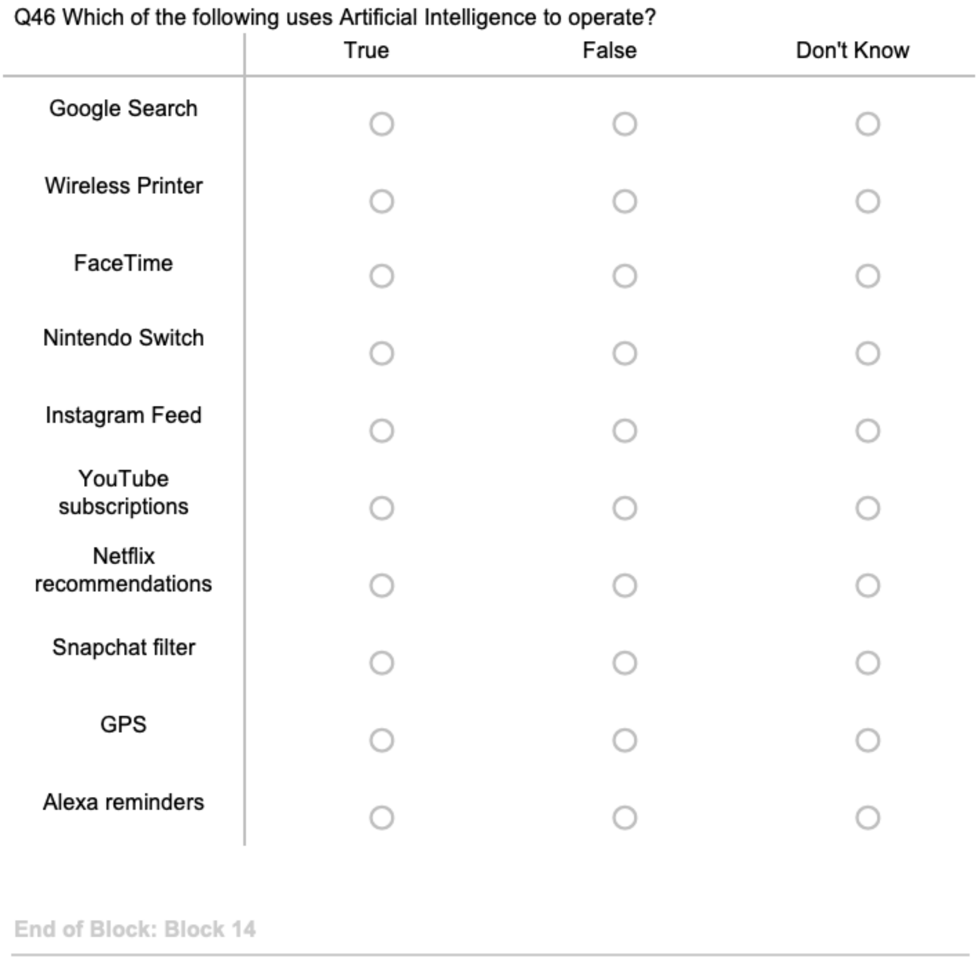
____

Q42 What groups of people are concerned with the outcomes of Google search engine? (Please separate each group with a comma)

____

End of Block: Block 10

Start of Block: Block 14



Q46 Which of the following uses Artificial Intelligence to operate?

	True	False	Don't Know
Google Search	○	○	○
Wireless Printer	○	○	○
FaceTime	○	○	○
Nintendo Switch	○	○	○
Instagram Feed	○	○	○
YouTube subscriptions	○	○	○
Netflix recommendations	○	○	○
Snapchat filter	○	○	○
GPS	○	○	○
Alexa reminders	○	○	○

End of Block: Block 14

B.2 End of Day Reflections

Two end of day written reflections were used throughout the workshop. They are presented below chronologically in the order that the students engaged with them.

B.2.1 Standard End of Day Reflection

This reflection was used on days 1-4 of the summer pilot workshop. Students were asked the following questions:

1. What did you learn today?
2. What did you add to the mural this afternoon?
3. What was your favorite activity today?
4. What activity did you learn the most from?

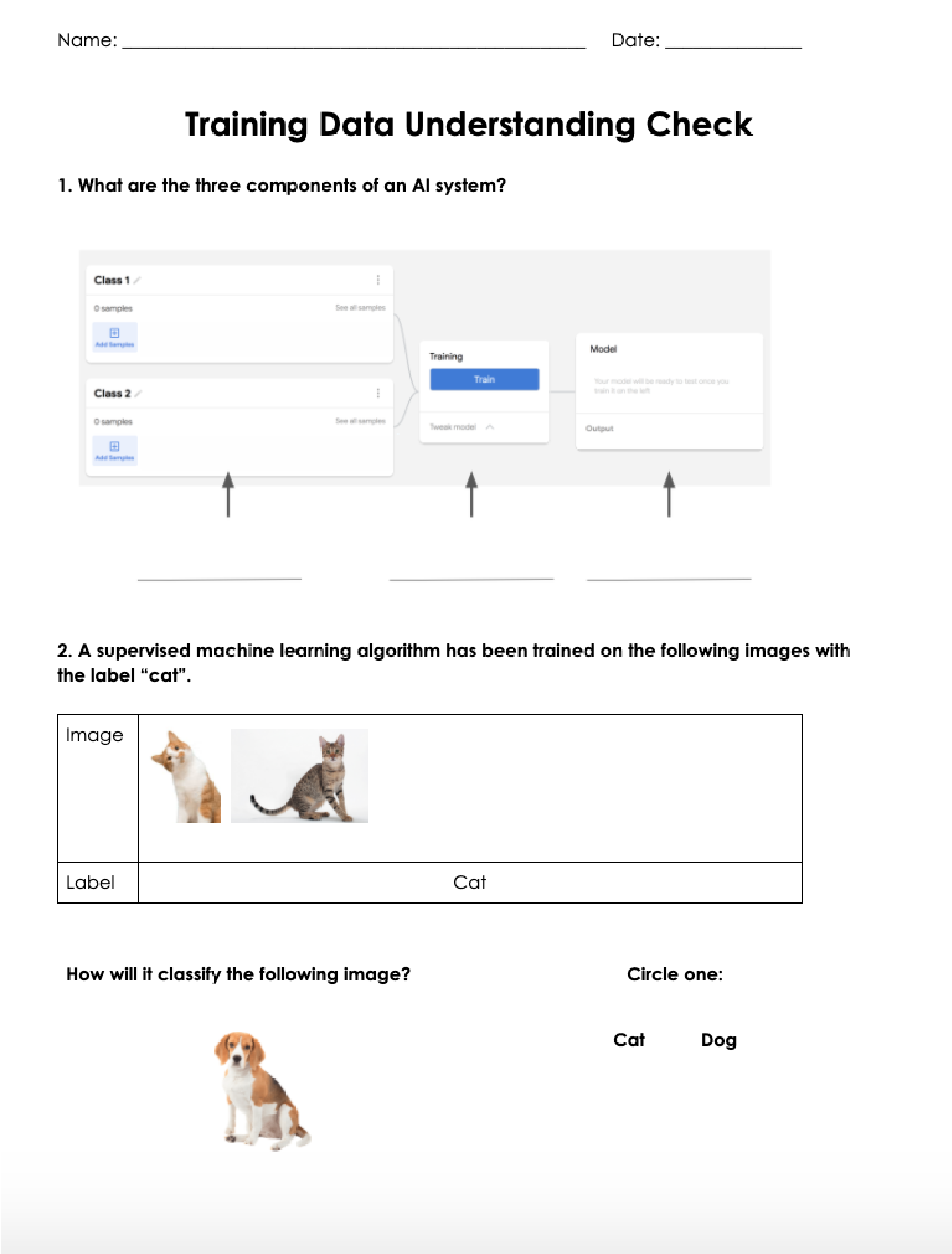
B.2.2 Final Reflection

On the last day of the workshop, students were asked to complete the following written reflection:

1. What was the 1 thing you learned all week?
2. What did you add to the mural today?
3. How has what you've added to the mural changed over time?
4. What was your favorite activity this week?

B.3 Teachable Machine Understanding Check

The following assessment was used to in conjunction with the Teachable Machines activity. Question 2 was based on [104].



Name: ______________________________ Date: ____________

Training Data Understanding Check

1. What are the three components of an AI system?

Class 1
0 samples
See all samples
Add Samples
Class 2
0 samples
See all samples
Add Samples
Training
Train
Tweak model
Model
Your model will be ready to test once you train it on the left
Output

______________ ______________ ______________

2. A supervised machine learning algorithm has been trained on the following images with the label "cat".

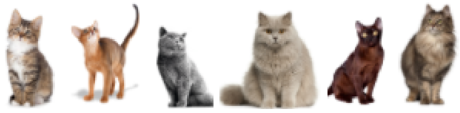
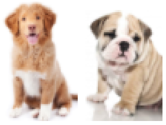
Image	
Label	Cat

How will it classify the following image?

Circle one:

Cat **Dog**

3. A supervised machine learning algorithm has been trained on the following images:

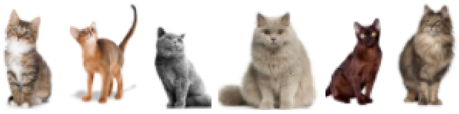
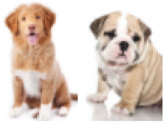
Image		
Label	Cat	Dog

3.1 Do you expect the algorithm's accuracy to be (circle one):

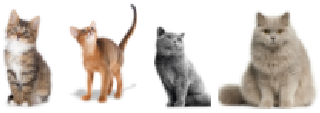
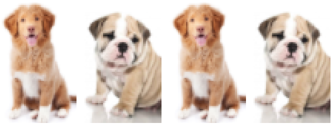
Better for cats **The same between cats and dogs** **Better for dogs**

3.2. Which training set, if any, would provide the best classification accuracy for both cats and dogs?

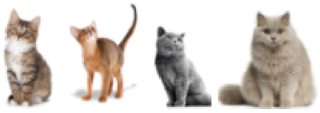
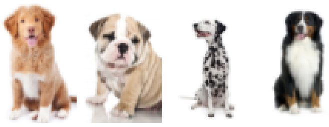
A.

	
Cat	Dog

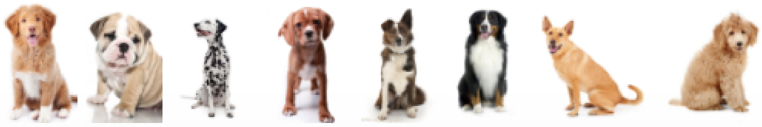
B.

	
Cat	Dog

C.

	
Cat	Dog

D.


Dog

B.4 Socratic Seminar Guided Questions

The following questions were used to guide the Socratic seminar activity. While these questions were pre-determined before the activity, students also presented their own questions to guide the conversation.

1. Can someone summarize this article?
2. What is the goal of this redesign? What is this platform optimizing for?
3. Can anyone name the stakeholders addressed in this article?
4. Who is the most important stakeholder?
5. Which stakeholder is making the most change or has the most power?
6. Do you think YouTube kids should be a separate product? Why? Why not?
7. Have you ever seen an inappropriate piece of content on YouTube? What did you do?
8. Would you use YouTube Kids app?
9. Do your parents make you use content controls now?
10. Would your parent like it [YouTube Kids app]?
11. Would a younger/older sibling like it?
12. How do you think advertisers feel about this?
13. Do you think it will be popular?
14. Do you think YouTube will lose profit? Is it okay if they lose profit?
15. What happens if there is more/less inappropriate content?
16. Why does autoplay exist? Who benefits from autoplay? Should autoplay exist?

C
Educational Resources

Below are the online educational resources used throughout the summer workshop (described in Chapters 8 and 9).

C.1 Teachable Machines

Two versions of Teachable Machine were used in the summer workshop.

Version 1, which is primarily a visual explainer and great for quick demonstrations, can be found at https://teachablemachine.withgoogle.com/v1/.

Version 2, what is now known as "Teachable Machine" can be found at https://teachablemachine.withgoogle.com/. This version allows users to train their own models, save them, and use them later.

C.2 Speculative Fiction Activity Tools

Students engaged with four different technologies during this activity, and groups were assigned a technology at random. The corresponding worksheets for this activity can be found at bit.ly/mit-ai-ethics.

GanPaint is an online, digital paint tool built on generative adversarial netoworks, a kind of AI. Instead of a typical digital paintbrush, the “brushes” in this tool can paint different styles - such as grass, sky, or doors - into a given image. The tool can be found at http://gandissect.res.ibm.com/ganpaint.html.

Talk to Transformer uses neural networks to generate text. Users can input some starter text and this tool will “complete” the text. This tool can be found at https://talktotransformer.com/.

Affectiva is software that predicts the emotional reactions of a user based on their facial expressions. The following demo gives an example of the software but does not store images of the user’s face. This demo can be found at https://demo.mr.affectiva.com/.

Deep Angel is another tool that utilizes generative adversarial networks to “paint” images. In this tool, however, the “brushes” can remove objects such as animals, peoples, or furniture from images. This tool can be found at http://deepangel.media.mit.edu/.

C.3 "Free Time" Activities and Resources

During students’ free time, many of them also engaged with the following two online resources.

Google Quick Draw is an online pictionary game where the user is given a word to draw and artificial intelligence software predicts what the given word is based upon the user’s drawing.

NVIDIA GauGAN is another GAN paint tool students used, which can be found at http://nvidia-research-mingyuliu.com/gaugan/.

D

Quotes from Women in AI, Ethics, and Education

The quotes appearing at the beginning of Chapters 1-11 in this thesis are all taken from amazing women who study, educate, and communicate about AI and ethics, many of whom I have had the privilege of knowing, and many of whom have shaped this work. Below I attribute each of the quotes used throughout this thesis and highlight the woman behind the words.

D.1 Tess Posner

Artificial intelligence is tech but also an idea. We're creating ourselves in a machine.

Tess Posner is the CEO of AI4ALL, a nonprofit program which seeks to make the field of artificial intelligence more inclusive and accessible, particularly to minority students. This quote was from a presentation she gave to AI4ALL, quoted by Sarah Eli Judd here: https://twitter.com/SarahEJudd/status/1224477730634457088?s=20. You can learn more about AI4ALL at http://ai-4-all.org/.

D.2 Cathy O'Neil

Algorithms are opinions embedded in code.

Cathy O'Neil is a mathematician and data scientist turned algorithmic auditor and advocate. Her book, "Weapons of Math Destruction" has had a tremendous impact in getting the tech industry, academia, and public at large to see algorithms as political artifacts. This quote is taken from her TED talk, found here: https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data _must_end. You can learn more about Cathy O'Neil's recent work at https://mathbabe.org/.

D.3 Casey Fiesler

...one of many steps that need to be taken toward this change is at the level of education - and whatever else we can do to ensure that no one is 'just an engineer' anymore.

Casey Fiesler is an assistant professor in the Department of Information Science at the University of Colorado Boulder where she studies governance in online communities, technology ethics, and fandom. In particular, she studies how to teach ethics to college students and has crowd-sourced and analyzed over 200 collegiate tech ethics syllabi. This quote is taken from an article titled "What Our Tech Ethics Crisis Says About the State of Computer Science," found at https://howwegettonext.com/what-our-tech-ethics-crisis-says-about-the-state-of-computer-science-education-a6a5544e1da6. You can also learn more about Fiesler's work at https://caseyfiesler.com/.

D.4 Randi Williams

> *...it is important that we democratize AI now. When anyone can learn about and use AI in creative ways, then AI can become a tool for positive change.*

Randi Williams is a PhD candidate in the Personal Robots Group at the MIT Media Lab. Her work concerns how we might educate children about artificial intelligence and equip them with the mindsets that they, too, can become engineers. This quote is from her masters thesis titled, "PopBots: Leveraging Social Robots to Aid Preschool Children's Artificial Intelligence Education." You can learn more about her work at https://www.media.mit.edu/people/randiw12/overview/.

D.5 Vivienne Ming

> *AI for Good is easy; it's AI-That's-Not-Bad that's hard.*

Vivienne Ming is the founder and executive chair of the think tank Socos Labs. She is a neuroscientist and artificial intelligence expert working to reframe the meaning of "AI for good." This quote is taken from a profile written by Kathy Baxter, titled "Inaugural Women in AI Ethics Summit" found at: https://medium.com/datadriveninvestor/inaugural-women-in-ai-ethics-summit-5440bd59da45. You can learn more about Ming's work at https://www.socos.me/vivienne.

D.6 Madeleine Clare Elish

Instead of saying '"deploy" technology, I prefer to use the word 'integrate.' Because it prompts the question '"into what?"

Madeleine Clare Elish is a cultural anthropologist at Data & Society, a research institution dedicated to studying the cultural implications of data-centric technologies and automation. Her work has shifted conversations about the ethics of AI away from fairness toward ethical design. This quote is provided by Kate Darling at https://twitter.com/grok_/status/1159195106102206464?s$\bar{2}$0. You can learn more about Elish's work at https://datasociety.net/people/elish-madeleine-clare/.

D.7 Joy Buolamwini

Whoever codes the system embeds her views. Limited views create limited systems. Let's code with a more expansive gaze.

Joy Buolamwini is a poet of code, PhD candidate at the MIT Media Lab, and founder of the Algorithmic Justice League. Her research investigates algorithmic bias, particularly in facial recognition software, and she advocates for greater responsibility and accountability by technology companies. This quote comes from an article she wrote titled, "InCoding – In the Beginning Was The Coded Gaze" which can be found at https://medium.com/mit-media-lab/incoding-in-the-beginning-4e2a5c51a45d. You can learn more about her work and the Algorithmic Justice League at https://www.ajlunited.org/.

D.8 Sherri Spelic

That's the kind of design thinking I hope and wish for: Where 'what's wrong' drives our pursuit of 'what if?'

Sherri Spelic is an educator, author, and communicator and the founder of the publication *Identity, Education, and Power.* She writes on the intersection of these three themes and on design in education. This quote is taken from her book *Care at the Core: Conversational Essays on Identity, Education, and Power.* You can read more of her writing at https://edifiedlistener.blog/.

D.9 Rachel Thomas

> *It is not only possible that everybody get involved with AI, it is actually crucial that you do so.*

Rachel Thomas is the director of the USF Center for Applied Data Ethics and co-founder of fast.ai, an organization dedicated to making deep neural network technology accessible to all. In addition to being a machine learning researcher and practitioner, Thomas's work also centers around making the field of AI more inclusive and diverse, especially to those who would not usually see themselves fitting into the field of AI. This quote is taken from her Tedx Talk, found at https://www.ted.com/talks/rachel_thomas_artificial_intelligence_needs_all_of_us. You can learn more about her work at https://www.fast.ai/topics/ai-in-society.

D.10 Hannah Fry

> *This is the future I'm hoping for.... One where we stop seeing machines as objective masters and start treating them as we would any other source of power.*

Hannah Fry is an associate professor in the Mathematics of Cities at the Centre for Advanced Spatial Analysis at University College London. Additionally, she is an author and mathematics communicator, with an emphasis on making mathematics and computer science accessible to the public. This quote is taken from her book titled *Hello, World: How to Be Human in the Age of the Algorithm.* You can learn more about her work at http://www.hannahfry.co.uk/.

D.11 Kate Crawford

> *Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations.*

Kate Crawford is the co-director and co-founder of the AI Now Institute at New York University, the world's first institution solely dedicated to studying the societal implications of artificial intelligence. Her research along with the work done at AI Now provides insight to the current state of "AI + Ethics" as it relates to research and measures for accountability. This quote comes from an article Crawford wrote in the *Harvard Business Review* titled "The Hidden Biases in Big Data," found here: https://hbr.org/2013/04/the-hidden-biases-in-big-data. You can learn more about Kate Crawford at https://www.katecrawford.net/.

Bibliography

[1] 2019 diversity report. Retrieved from https://diversity.fb.com/read-report/ on 8-25-2019.

[2] AI experiments. Retrieved from https://experiments.withgoogle.com/collection/ai on 3-15-2020.

[3] AI for all: Summer programs. Retrieved from http://ai-4-all.org/summer-programs/ on 3-15-2020.

[4] AI for good: About us. Retrieved from https://aiforgood.itu.int/about-us/ on 3-15-2020.

[5] AI Topics. Retrieved from https://aitopics.org/search on 1-15-2020.

[6] Banning the usage of facial recognition technology in Somerville. Retrieved from http://somervillecityma.iqm2.com/Citizens/Detail_LegiFile.aspx?ID¯20991& highlightTerms¯facial%20recognition%20technology on 3-15-2020.

[7] Be Internet Awesome. Retrieved from https://beinternetawesome.withgoogle.com/en_us/ on 3-15-2020.

[8] A Brief History of AI, note = Retrieved from https://aitopics.org/misc/brief-history on 1-15-2020.

[9] Calypso for Cozmo. Retrieved from https://calypso.software/ on 3-15-2020.

[10] Cognimates. Retrieved from http://cognimates.me/projects/ on 3-15-2020.

[11] eCraft2Learn. Retrieved from https://ecraft2learn.github.io/ai/ on 3-15-2020.

[12] EthicalCS: Inclusive, interdisciplinary computer science education. Retrieved from https://ethicalcs.org/ on 3-15-2020.

[13] Everything you need to teach digital citizenship. Retrieved from https://www.commonsense.org/education/digital-citizenship on 3-15-2020.

[14] Facial recognition technology (Part II): Ensuring transparency in government use. Retrieved from https://oversight.house.gov/legislation/hearings/facial-recognition-technology-part-ii-ensuring-transparency-in-government-use on 3-15-2020.

[15] Gobo. Retrieved from https://gobo.social/ on 3-05-2019.

[16] Machine learning for kids. Retrieved from https://machinelearningforkids.co.uk/#!/worksheets on 3-15-2020.

[17] Machine Learning: Overviews. Retrieved from https://aitopics.org/search?filters=taxnodes%3ATechnology%7CInformation +Technology%7CArtificial+Intelligence%7CMachine+Learning%40%40taxnodes %3AGenre%7COverview on 1-15-2020.

[18] News and america's kids: How young people perceive and are impacted by the news. Technical report, Common Sense Media.

[19] NVIDIA techsplorer. Retrieved from https://www.nvidia.com/en-us/foundation/programs/techsplorer-stem-program/ on 3-15-2020.

[20] Technovation families AI lessons. Retrieved from https://www.curiositymachine.org/lessons/lesson/ on 3-15-2020.

[21] Teens in AI: About us. Retrieved from https://www.teensinai.com/#about_us on 3-15-2020.

[22] Tensorflow playground, note = "retrieved from https://playground.tensorflow.org/ on 3-15-2020".

[23] Use the "pb&j" sandwich activity to introduce important components of algorithms. Retrieved from https://www.csteachingtips.org/tip/use-pbj-sandwich-activity-introduce-important-components-algorithms on 3-05-2019.

[24] What if... Retrieved from https://pair-code.github.io/what-if-tool/ on 3-15-2020.

[25] World artificial intelligence competition for youth. Retrieved from https://www.waicy.org/waicy-2020/ on 3-15-2020.

[26] Aaron Smith, Skye Toor, and Patrick Van Kessel. Many turn to YouTube for children's content, news, how-to lessons. Technical report, Pew Research Center.

[27] Justin Aglio. Coming this fall to montour school district: America's first public school AI program. Retrieved from https://www.gettingsmart.com/2018/07/coming-this-fall-to-montour-school-district-americas-first-public-school-ai-program/ on 3-05-2019.

[28] Monica Anderson and Jingjing Jiang. Teens, social media technology 2018. Technical report, Pew Research Center.

[29] Eric P.S. Baumer, Timothy Berrill, Sarah C. Botwinick, Jonathan L. Gonzales, Kevin Ho, Allison Kundrik, Luke Kwon, Tim LaRowe, Chanh P. Nguyen, Fredy Ramirez, and et al. What would you do? design fiction and ethics. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, GROUP '18, page 244–256, New York, NY, USA, 2018. Association for Computing Machinery.

[30] Timothy Bell, Jason Alexander, Isaac Freeman, and Mick Grimley. Computer science unplugged: school students doing real computing without computers. *The New Zealand Journal of Applied Computing and Information Technology*, 13, 01 2009.

[31] Debra Bernstein and Kevin Crowley. Searching for signs of intelligent life: An investigation of young children's beliefs about robot intelligence. *Journal of The Learning Sciences - J LEARN SCI*, 17:225 – 247, 04 2008.

[32] John Biggs. Programmer creates an AI to (not quite) beat NES games. Retrieved from https://techcrunch.com/2013/04/14/nes-robot/ on 3-15-2020.

[33] Dipan Bose, Maria Segui-Gomez, ScD, and Jeff R. Crandall. Vulnerability of female drivers involved in motor vehicle crashes: An analysis of us population at risk. *American Journal of Public Health*, 101(12):2368–2373, 2011. PMID: 22021321.

[34] Michael Buckley, John Nordlinger, and Devika Subramanian. Socially relevant computing. *ACM SIGCSE Bulletin*, 40(1):347 – 351, 2008.

[35] Joy Buolamwini. Gender shades. Retrieved from https://www.youtube.com/watch?v=TWWsW1w-BVot=7s on 1-22-2019.

[36] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[37] Adam Burke. Occluded algorithms. *Big Data & Society*, 6(2):2053951719858743, 2019.

[38] Emanuelle Burton, Judy Goldsmith, and Nicholas Mattei. How to teach computer ethics through science fiction. *Commun. ACM*, 61(8):54–64, July 2018.

[39] Mary Elaine Califf and Mary Goodwin. Effective incorporation of ethics into courses that focus on programming. *SIGCSE Bull.*, 37(1):347–351, February 2005.

[40] Stephen Cave, Kate Coughlan, and Kanta Dihal. "scary robots": Examining public responses to AI. pages 331–337, 01 2019.

[41] Sasha Costanza-Chock. Design justice, AI, and escape from the matrix of domination. 2018.

[42] Sasha Costanza-Chock. Design justice: towards an intersectional feminist framework for design theory and practice. *Proceedings of the Design Research Society*, 2018.

[43] danah boyd. You think you want media literacy... do you? Retrieved from https://points.datasociety.net/you-think-you-want-media-literacy-do-you-7cad6af18ec2 on 2-5-2020.

[44] Jeffery Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G on 3-15-2020.

[45] Fred Martin David S. Touretzky, Christina Gardner-McCune and Deborah Seehorn. K-12 guidelines for artificial intelligence: What students should know. Retrieved from https://github.com/touretzkyds/ai4k12/raw/master/documents/ISTE_2019 _Presentation_website_final.pdf on 3-24-2019.

[46] Janet Davis and Henry M. Walker. Incorporating social issues of computing in a small, liberal arts college: A case study. In *Proceedings of the 42Nd ACM Technical Symposium on Computer Science Education*, SIGCSE '11, pages 69–74, New York, NY, USA, 2011. ACM.

[47] Janet Davis and Henry M. Walker. Incorporating social issues of computing in a small, liberal arts college: A case study. In *Proceedings of the 42nd ACM Technical Symposium on Computer Science Education*, SIGCSE '11, page 69–74, New York, NY, USA, 2011. Association for Computing Machinery.

[48] Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. "Hey Google is it OK if i eat you?": Initial explorations in child-agent interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children*, IDC '17, pages 595 – 600, New York, NY, USA, 2017. Association for Computing Machinery.

[49] Arielle Duhaime-Ross. Apple promised an expansive health app, so why can't i track menstruation? Retrieved from https://www.theverge.com/2014/9/25/6844021/apple-promised-an-expansive-health-app-so-why-cant-i-track on 1-20-2020.

[50] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. "I always assumed that I wasn't really that close to her": Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 153 – 162, New York, NY, USA, 2015. Association for Computing Machinery.

[51] Casey Fiesler. Black mirror, light mirror: Teaching technology ethics through speculation. Retrieved from https://howwegettonext.com/the-black-mirror-writers-room-teaching-technology-ethics-through-speculation-f1a9e2deccf4 on 1-22-2019.

[52] Casey Fiesler. What our tech ethics crisis says about the state of computer science education. *How We Get to Next*, 2018.

[53] Casey Fiesler, Natalie Garrett, and Nathan Beard. What do we teach when we teach tech ethics? a syllabi analysis. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, SIGCSE '20, page 289–295, New York, NY, USA, 2020. Association for Computing Machinery.

[54] Douglas Fisher and Nancy Frey. Gradual release of responsibility instructional framework. *IRA E-ssentials*, pages 1–8, 2013.

[55] Max Fisher and Amanda Taub. How youtube radicalized brazil. August 2019. Retrieved from https://www.nytimes.com/2019/08/11/world/americas/youtube-brazil.html?searchResultPosition=1 on 3-16-2020.

[56] Batya Friedman. Value-sensitive design. *interactions*, 3(6):16–23, 1996.

[57] Hannah Fry. *Hello, World: Being Human in the Age of Algorithms*, chapter Power, pages 14–15. W. W. Norton & Company, 2018.

[58] Natalie Garrett, Nathan Beard, and Casey Fiesler. More than "if time allows": The role of ethics in AI education. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 272–278, New York, NY, USA, 2020. Association for Computing Machinery.

[59] Rachel Goodman. Why Amazon's automated hiring tool discriminated against women. Retrieved from https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against on 1-22-2019.

[60] Barbara J Grosz, David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo. Embedded EthiCS: integrating ethics across CS education. *Communications of the ACM*, 62(8):54–61, 2019.

[61] Cathy Gun, Hanna O'Neil. Near term ai. *Ethics of Artificial Intelligence.*

[62] Tristan Harris. Tristan [h]arris's medium. Retrieved from https://medium.com/@tristanharris on 3-22-2020.

[63] Matthew Hutson. Artificial intelligence could identify gang crimes—and ignite an ethical firestorm. *Science*, pages 23–25, 02 2018.

[64] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhu Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Technical report, Harvard University.

[65] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. Machine bias. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing on 1-20-2020.

[66] Ken Kahn. A guide to building AI apps and artefacts: Chapter 3 - Adding image recognition to programs. Retrieved from https://ecraft2learn.github.io/ai/AI-Teacher-Guide/chapter-3.html on 3-15-2020.

[67] Rimma Kats. How much do people know about AI? May 2017. Retrieved from https://www.emarketer.com/Article/How-Much-Do-People-Know-About-AI/1015949 on 3-15-2020.

[68] Nazish Zaman Khan and Andrew Luxton-Reilly. Is computing for social good the solution to closing the gender gap in computer science? In *Proceedings of the Australasian Computer Science Week Multiconference*, ACSW '16, New York, NY, USA, 2016. Association for Computing Machinery.

[69] Lawrence Kohlberg. *The philosophy of moral development: Moral stages and the idea of justice.* Harper & Row, 1981.

[70] Dale Lane. Intriguing student ml project - looking to see if an ml model can recognize if a child is being bullied, based on the pictures they draw. obviously needs care about how such a system would be used, but i wonder if technically it would work? Tweeted 5-11-2019.

[71] Mark Latonero. Opinion: AI for good is often bad. Retrieved from https://www.wired.com/story/opinion-ai-for-good-is-often-bad/ on 3-15-2020.

[72] Sharona T. Levy and David Mioduser. Approaching complexity through planful play: Kindergarten children's strategies in constructing an autonomous robot's behavior. *International Journal of Computers for Mathematical Learning*, 15:21–43, 2010.

[73] Lindsay-Rae McIntyre. Diversity and inclusion update: The journey continues. Retrieved from https://blogs.microsoft.com/blog/2018/11/14/diversity-and-inclusion-update-the-journey-continues/ on 8-25-2019.

[74] Emily McReynolds, Sarah Hubbard, Timothy Lau, Aditya Saraf, Maya Cakmak, and Franziska Roesner. Toys that listen: A study of parents, children, and internet-connected toys. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 5197–5207, New York, NY, USA, 2017. Association for Computing Machinery.

[75] Karen G. Mills. Gender bias complaints against apple card signal a dark side to fintech. Retrieved from https://hbswk.hbs.edu/item/gender-bias-complaints-against-apple-card-signal-a-dark-side-to-fintech on 1-20-2020.

[76] Arvind Narayanan and Shannon Vallor. Why software engineering courses should include ethics coverage. *Communications of the ACM*, 57:23–25, 03 2014.

[77] Jack Nicas. How youtube drives people to the internet's darkest corners. February 2018. Retrieved from https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478 on 3-16-2020.

[78] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press, 2018.

[79] Elements of AI. Course overview. Retrieved from https://course.elementsofai.com/ on 3-15-2020.

[80] Cathy O'Neil. Know thy futurist. Retrieved from http://bostonreview.net/science-nature-podcast/cathy-oneil-know-thy-futurist on 1-20-2020.

[81] Melonie Parker. Google diversity annual report 2019.

[82] Caroline Criado Perez. *Invisible Women: Data Bias in A World Designed for Men.* Harry N. Abrams, 2019.

[83] ReadyAI. About us. Retrieved from https://www.readyai.org/about/ on 3-15-2020.

[84] Kevin Roose. The making of a youtube radical. June 2019. Retrieved from https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html on 3-16-2020.

[85] Ryan Saavedra. Socialist rep. alexandria ocasio-cortez (d-ny) claims that algorithms, which are driven by math, are racist. Tweeted 1-22-2019.

[86] Doris Schroeder and Clare Palmer. Technology assessment and the'ethical matrix'. *Poiesis & Praxis*, 1(4):295–307, 2003.

[87] Exploring Computer Science. Artificial intelligence: Alternate curriculum unit. Retrieved from http://www.exploringcs.org/for-teachers-districts/artificial-intelligence on 3-15-2020.

[88] Nick Seaver. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, 4(2):2053951717738104, 2017.

[89] Rachel L. Severson and Stephanie M Carlson. Behaving as or behaving as if? children's conceptions of personified robots and the emergence of a new ontological category. October 2010.

[90] Katie Shilton. Values and ethics in human-computer interaction. *Found. Trends Hum.-Comput. Interact.*, 12(2):107–171, July 2018.

[91] Tom Simonite. AI is the future—but where are the women? Retrieved from https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/ on 8-25-2019.

[92] Michael Skirpan, Nathan Beard, Srinjita Bhaduri, Casey Fiesler, and Tom Yeh. Ethics education in context: A case study of novel ethics activities for the CS classroom. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, SIGCSE '18, page 940–945, New York, NY, USA, 2018. Association for Computing Machinery.

[93] Jackie Snow. Bias already exists in search engine results, and it's only going to get worse. Retrieved from https://www.technologyreview.com/s/610275/meet-the-woman-who-searches-out-search-engines-bias-against-women-and-minorities/ on 1-20-2020.

[94] Shane Snow. How soylent and oculus could fix the prison system (a thought experiment). Retrieved from https://medium.com/@shanesnow/how-soylent-and-oculus-could-fix-the-prison-system-a-thought-experiment-e26be8b21a42 on 3-05-2019.

[95] Carol Spradling, Leen-Kiat Soh, and Charles Ansorge. Ethics training and decision-making: Do computer science programs need help? In *Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education*, SIGCSE '08, pages 153–157, New York, NY, USA, 2008. ACM.

[96] About Amazon Staff. Our workforce data. Retrieved from https://www.aboutamazon.com/working-at-amazon/diversity-and-inclusion/our-workforce-data on 8-25-2019.

[97] Tom Terrific. Kids & tech: The evolution of today's digital natives. Technical report, Influence Central, 2019.

[98] Rachel Thomas. "diversity and ai ethics are not separate issues. you can't have ethics without diversity." miad #womeninai. Tweeted 8-21-2019.

[99] Till Alexander Leopold , Saadia Zahidi, and Vesselina Ratcheva. The future of jobs report. Technical report, World Economic Forum.

[100] Ariana Tobin. Hud sues facebook over housing discrimination and says the company's algorithms have made the problem worse. Retrieved from https://www.propublica.org/article/hud-sues-facebook-housing-discrimination-advertising-algorithms on 3-15-2020.

[101] David Touretzky, Christina Gardner-McCune, Fred Martin, and Deborah Seehorn. Envisioning AI for K-12: What should every child know about AI? *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:9795–9799, 07 2019.

[102] Zeynep Tufekci. YouTube, the great radicalizer. March 2018. Retrieved from https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html on 3-16-2020.

[103] James Vincent. Tencent says there are only 300,000 AI engineers worldwide, but millions are needed. December 2017. Retrieved from https://www.theverge.com/2017/12/5/16737224/global-ai-talent-shortfall-tencent-report on 1-15-2020.

[104] Randi Williams. PopBots: Leveraging social robots to aid preschool children's artificial intelligence education. Master's thesis, Massachusetts Institute of Technology, Media Lab, June 2018.

[105] Randi Williams, Christian Vázquez Machado, Stefania Druga, Cynthia Breazeal, and Pattie Maes. "My doll says it's ok": A study of children's conformity to a talking doll. In *Proceedings of the 17th ACM Conference on Interaction Design and Children*, IDC '18, pages 625 – 631, New York, NY, USA, 2018. Association for Computing Machinery.

[106] Langdon Winner. *Do Artifacts Have Politics?*, volume 109, pages 26–38. 01 1985.

[107] Sarah Wu. Somerville City Council passes facial recognition ban. Retrieved from https://www.bostonglobe.com/metro/2019/06/27/somerville-city-council-passes-facial-recognition-ban/SfaqQ7mG3DGulXonBHSCYK/story.html on 3-15-2020.