AI Ethics Curriculum Development for Researchers and Future Policy-Makers

Benjamin Boudreaux, Jarrett Catlin, Sarah Denton, Osonde Osoba, Tepring Piquado, Patricia Stapleton

The RAND Corporation and Pardee RAND Graduate School, September 2020

This project involved three major tasks:

(1) Analyzing existing Artificial Intelligence (AI) university courses to identify gaps and key barriers to integrating ethics into the AI curriculum, with a specific focus on the Pardee RAND Graduate School (full analysis attached as Appendix 1)

(2) Developing an annotated syllabus for a modular AI ethics curriculum (annotated syllabus attached as Appendix 2)

(3) Organizing an 'ethics hackathon' involving participants at the Pardee RAND Graduate School in collaboration with the non-profit COVID Alliance to foster ethical deliberation and actionable recommendations on a real-world problem (hackathon overview attached as Appendix 3)

1. Summary of key findings

- There is a burgeoning effort to integrate ethics into AI coursework, but there are gaps related to several key ethical issues, including diversity in AI, accessibility, the human element in AI development, environmental issues related to AI, and military AI applications beyond autonomous weapons.
- Of 32 courses reviewed at Pardee RAND, approximately 19% integrated AI ethics topics. There are opportunities for further embedding ethical topics into coursework through increased collaboration among professors and students.
- The annotated syllabus provides a modular approach to an AI ethics curriculum, and includes a set of readings and discussion questions to be used to further integrate AI ethics topics or to administer AI ethics courses.
- The ethics hackathon concept is a valuable approach to enable student's ethical thinking related to AI and to develop actionable proposals for AI technology.

2. Background and Problem Definition

a. What was the project's main objective?

The main objective was to identify opportunities to integrate ethical thinking into AI related coursework and research at the Pardee RAND Graduate School, and to facilitate discussion of AI ethics. Ultimately our goal was to enable the students, many of whom will go on to be AI researchers and policy-makers, to consider the ethical implications of AI technology and to practice developing risk mitigations. This objective is aligned to the broader goal at Pardee RAND to integrate ethics as a cross-cutting thread across its policy engagement streams.

b. What was the initial problem you wanted to solve?

In many cases, AI technology is researched, developed, and brought to market without reflection on the potential harms or risks that the technology might entail. AI technology researchers and developers are not normally trained to reflect on the ways their research or products might be misused or abused, or on the long-term implications of AI, especially for vulnerable and marginalized populations. In addition, these technologists are often from homogenous demographic groups and do not include diverse perspectives that would help identify potential risks. This is a systemic problem that we have sought to begin to address by identifying opportunities in the curriculum to inject ethics and to execute an activity that required ethical reflection related to AI.

c. Who/what are other individuals or institutions working on similar projects?

AI ethics is a rapidly developing research area that universities, technology companies, and other organizations have sought to address. Several important and influential organizations are described in the annotated syllabus, including the Algorithmic Justice League and the AI Now Institute. The syllabus also has links to dozens of important papers that address AI ethics issues from scholars and researchers worldwide. In addition, our work evaluating ethics issues and gaps in existing coursework drew from research and a dataset compiled by researchers at the University of Colorado, Boulder. Our list of these organizations and academics involved in AI ethics work is just a sampling of the many groups seeking to make progress on the issue.

d. *Did you work with other teams or institutions? If yes, how?*

For the ethics hackathon, we partnered with the COVID Alliance, a non-profit coalition focused on developing a coordinated response to COVID-19, consisting of epidemiologists, virologists, policy experts, medical informaticians, data scientists, and software engineers. In particular, we worked with the COVID Alliance's research management platform which is intended to aggregate and analyze data to be used for research that assists in COVID response. There is more detail on the partnership in the write-up describing the ethics hackathon (attached as Appendix 3).

e. *How did you define diversity, equity and inclusion with respect to your work?*

We defined diversity and inclusion in several ways. First, we sought out a diverse team to develop the analysis and take part in the hackathon. This included RAND researchers, Pardee RAND professors, and Pardee RAND students with a variety of backgrounds, including racial and ethnic minorities, international students, and various levels of seniority and research interests across the RAND Corporation. Second, we directly discuss diversity related challenges in the development and fielding of AI in the course analysis memo, and the topic of diversity in AI development is a key issue we highlight in the annotated syllabus—please see those documents for additional detail. Third, the primary focus of the ethics hackathon was to explore the potential implications for equity and harms to minority and marginalized communities in the use of the COVID Alliance research platform. Students participating in the hackathon were explicitly asked to assess the implications for equity and to recommend potential solutions.

3. *Development*

a. *How did you first approach the project? i. What were the intended methods and processes you wanted to use?*

We first approached the project with the goal of convening several discussions involving the Pardee RAND community to discuss specific issues related to AI ethics. However, the move to a remote environment somewhat stalled our plans for these in-person discussions. In addition, the co-lead of the project took a leave of absence from RAND which required some shifting of responsibility.

b. *What changes did you make to the project? i. How did you adapt to any changes in circumstances for the project?*

The major change we made to the project was to take advantage of a partnership opportunity with the COVID Alliance to anchor the ethical reflection activities we had planned to inform a real existing AI technology. Several Pardee RAND graduate students and professors had built a relationship with the COVID Alliance, and we identified and pursued the chance to partner with the organization to explore the ethics of their research tool. We decided to develop and implement the ethics hackathon concept as a way to foster interdisciplinary discussion and team-based competition among graduate students. More details on the hackathon are in Appendix 3.

c. *How did you evaluate the success of the project?*

We have evaluated success of the project by the number of students engaged in the hackathon and the richness and actionability of their proposed recommendations. Other elements of the project, such as the further integration of AI ethics into the curriculum in specific courses, will need to be continuously evaluated over the coming months as discussions with existing professors proceed.

[redacted].

[redacted]

c. *Did you encounter diversity, equity, and inclusion challenges in your project? i. How did you respond to them?*

We encountered diversity, equity, and inclusion challenges both in terms of the methods of conducting the tasks and the content of the project. More details on the diversity and equity challenges associated with AI coursework and the COVID-Alliance platform are discussed in the attachments.

5. Lessons learned

a. *How would you summarize your insights?*

Per our summary of key findings above, here are some of our insights:

- There is a burgeoning effort to integrate ethics into AI coursework, but there are gaps related to several key ethical issues, including diversity in AI, accessibility, the human element in AI development, environmental issues related to AI, and military AI applications beyond autonomous weapons.
- Of 32 courses reviewed at Pardee RAND, approximately 19% integrated AI ethics topics. There are opportunities for further embedding ethical topics into coursework through increased collaboration among professors and students.
- The annotated syllabus provides a foundation for a modular curriculum on AI ethics, including a set of readings and discussion questions to be used to further integrate AI ethics topics or to administer AI ethics courses.
- The ethics hackathon concept is a valuable approach to enable student's ethical thinking related to AI and to develop actionable proposals for AI technology.

b. *What specific advice would you offer to other members with regards to this project?*

AI and technology ethics is a rapidly growing field, and there is important existing work happening in academic institutions, non-profits, and within technology companies. Someone looking to integrate AI ethics into a curriculum need not reinvent the wheel and can draw from the wealth of existing papers and efforts. The annotated syllabus is intended to build on these existing efforts and provide an easy-to-use description and set of resources of key AI ethics topics, however it is not comprehensive and is only a snap-shot of a rapidly developing field. We suggest that the syllabus be supplemented with additional resources tailored for their purpose.

6. *Possibilities to replicate*

1. *How can other members replicate the project, or part of the project?*
2. *What considerations should other members have when approaching your challenge?*

The project can be replicated in several ways. First, others can use the framework we present in the analysis memo (appendix 1) for evaluating the ethical issues within existing coursework at their institution. This approach helped to identify key gaps to be filled. Second, other institutions can use the annotated syllabus to build a standalone AI ethics course, or integrate AI ethics topics into other coursework. Third, other institutions can adopt and further develop the ethics hackathon concept described more thoroughly in appendix 3.

7. *General Information*

   1. *Who can be contacted to get more information?*

Benjamin Boudreaux, bboudrea@rand.org, 310.393.0411 x6197

   2. *What is the current state of the project?*

The project is complete.

8. *Annexes & Publications*

(1) AI Ethics Memo
(2) AI Ethics Annotated Curriculum
(3) AI Ethics Hackathon Overview

Annex 1

PROJECT MEMORANDUM

From: Sarah W. Denton, Benjamin Boudreaux, Patricia Stapleton

Date: September 30, 2020

Subject: AI Ethics for Researchers & Future Policy Makers

AI Ethics Curricula Overview

Policymakers, researchers, and members of the public have raised concerns about the ethics of artificial intelligence (AI).[1] These include concerns about the fairness of algorithms in high-stake social applications such as criminal justice, concerns about AI-enabled facial recognition infringing on personal privacy, and the trustworthiness and reliability of AI systems especially in military contexts. While these concerns are not new,[2] technical professionals are increasingly being called upon to exercise their responsibility to design, develop, and deploy AI and other emerging technologies in ways that benefit humanity.[3] To meet the increasing demand for technical talent, universities have scaled-up their AI and machine learning (ML) course offerings, but a critical question remains – how is ethics being incorporated into technical coursework?

[1] Skirpan, M., et al. (2018). "Ethics Education in Context: A Case Study of Novel Ethics Activities for the CS Classroom." *SIGCSE Proceedings of the 49th ACM Technical Symposium on Computer Science Education*; February: 940-945. https://dl.acm.org/doi/pdf/10.1145/3159450.3159573. Grosz, B.J., et al. (2019). "Embedded EthiCS: Integrating Ethics Across CS Education." *Communications of the ACM*; August 62(8): 54-61. https://cacm.acm.org/magazines/2019/8/238345-embedded-ethics/fulltext. Saltz, J., et al. (2019). "Integrating Ethics within Machine-Learning Courses." *ACM Trans. Comput. Educ.* 19(4). https://dl.acm.org/doi/pdf/10.1145/3341164.

[2] Nielsen, N.R. (1972). "Social responsibility and computer education." *SIGCSE: Proceedings of the 2nd SIGCSE technical symposium on Education in computer science*; March: 90-96. https://dl.acm.org/doi/pdf/10.1145/800155.805011.

[3] Data & Society researchers conducted an ethnographic study and interviewed 17 individuals working at well-known technology companies that are working to embed ethics in technology development. For more information, see, Metcalf, J., Moss, E., and Boyd, D. (2019). "Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics." *Data & Society* originally appeared in *Social Research: An International Quarterly*; 82(2): 449-476. https://datasociety.net/wp-content/uploads/2019/09/Owning-Ethics-PDF-version-2.pdf. Also see: Bughin, J., et al. (2019). "Tech for Good." *McKinsey Global Institute*; May. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/future%20of%20organizations/tech%20for%20good%20using%20technology%20to%20smooth%20disruption%20and%20improve%20well%20being/tech-for-good-mgi-discussion-paper.ashx.; Moss, E. and Metcalf, J. (2019). "The Ethical Dilemma at the Heart of Big Tech Companies." *Harvard Business Review*; 14 November. https://hbr.org/2019/11/the-ethical-dilemma-at-the-heart-of-big-tech-companies.

This memo presents an overview of our meta-analysis of AI Ethics curricula and provides insight into the integration level of ethics in technical courses as well as the ethical topics most often included in syllabi.

Analysis of AI Ethics Syllabi

We analyzed data collected by external researchers and supplemented this data with our own collection effort. Particular attention was paid to qualitative analyses conducted by researchers at the University of Colorado Boulder, which analyze the inclusion of ethics in 202 "tech ethics" courses[4] and 186 computer science (CS) courses[5] more broadly. After combining the datasets from the Garrett et al. and Saltz et al. studies, removing duplicate data inputs, and narrowing the focus to AI and ML specific courses, Garrett et al. found that only 51 courses incorporated ethics in the syllabi.[6] Of those 51 courses, 34 universities were represented.[7] Table 1 shows how these courses split between standalone AI ethics and technical courses.

Table 1: Breakdown of AI Ethics Courses

	# of Universities	Total Courses
	22 U.S. Universities	*31 Standalone AI Ethics Courses*
	12 U.S. Universities	*20 Technical Courses **
TOTAL	***34 U.S. Universities***	***51 Courses w/ Ethical Components***

***NOTE:** Out of nearly 200 AI/ML/Data Science courses, only 12% of technical courses included some mention of an ethics-related topic -- Garrett et al. analyzed 20 of the courses that did include an ethics component.

Tables 2-4 also draw upon the Garrett et al. combined dataset, depicting the ethical topics most often covered in standalone AI Ethics courses, topics covered most often in technical courses, and important ethical topics not represented in either technical or non-technical curricula. Table 2 breaks down the ethical topics covered in 31 standalone AI ethics courses by how many times a given topic was represented in the syllabi. Interestingly, privacy is only covered independently of other ethical topics like bias, policy and regulation, and consequences of algorithms in 26% of standalone AI Ethics syllabi. However, privacy is likely the most embedded since it spans across multiple ethical topics areas. The two topics most often covered are by far bias/fairness and

[4] Fiesler, C., Garrett, N., and Beard, N. (2020). "What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis." *SIGCSE*; March. https://cmci.colorado.edu/~cafi5706/SIGCSE2020_EthicsSyllabi.pdf.

[5] Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., Dewar, N., and Beard, N. (2019). "Integrating Ethics within Machine Learning Courses." *ACM Trans. Comput. Educ.*, 19(4) Article 32. https://dl.acm.org/doi/fullHtml/10.1145/3341164.

[6] Garrett, N., Beard, N., and Fiesler, C. (2020). "More Than 'If Time Allows': The Role of Ethics in AI Education." AIES; February: p. 3. https://cmci.colorado.edu/~cafi5706/AIES_EthicsEducation.pdf.

[7] According to a dataset provided by Casey Fiesler, one of the co-authors of Garrett et al., 90 universities worldwide are represented in the total dataset before removing duplicate inputs.

automation/robotics, which are also topics that receive high levels of public attention through popular media sources. Unsurprisingly, ethical theory ranks the lowest, with only 6% of AI Ethics syllabi including traditional philosophy readings.

Table 2: Courses by AI Ethical Topic Area

31 TOTAL STANDALONE AI ETHICS COURSES		
Topics	**% Of Syllabi**	***Examples***
Bias	87%	COMPAS recidivism algorithm / FR for sexuality prediction / Google mislabeling African American woman as a gorilla
Automation + Robotics	71%	Future of Work / LAWS / Self-driving cars / Robot rights / Future of Industries
Policy & Governance	55%	GDPR / Predictive policing
Philosophy & Morality	45%	Human responsibility + dignity / Morality generally / Incorporating morality into AI design / Existential threat of AI
Consequences of Algorithms	45%	Society-level consequences (e.g., impact of algorithms on democracy, civil & human rights, filter bubbles, targeted ads, etc.)
Privacy *****	32%	Cambridge Analytica / Surveillance / Big data
Future of AI	26%	General AI / Superintelligence / Forecasting Future AI tech / Future of Work / Singularity / Debates between Zuckerburg & Musk
History of AI	19%	Origins of computing / Turing / The Rise of Ai / Theory + philosophical concepts of intelligence
Ethical Theory	6%%	Utilitarianism / Deontology

*** NOTE:** Privacy appeared in conjunction with other topics like bias, regulation, and automation -- Therefore, its level of inclusion in AI Ethics syllabi is likely more pronounced than when considered as an isolated topic.

Table 3 breaks down the ethical topics covered by technical courses with ethical components. The ethical topics included in technical syllabi mirror those most often covered in standalone AI Ethics courses. Unfortunately, neither the Saltz et al. dataset nor the combined Garrett et al. dataset included information about the frequency of ethical topic inclusion in technical curricula.

Table 3: Technical Courses by AI Ethical Topic Area

20 TECHNICAL COURSES W/ ETHICS COMPONENT(S)	
Topics	***Examples***
Bias	*Avoiding* Bias

Fairness	*Promoting* Fairness
Privacy	*Protecting* Privacy / Differential privacy (ML framework used to. Mitigate risk of exposing sensitive data)
General Ethics Topics	With the exception of 1 course, the majority of courses covered ethics-related topics in the last two days of class "if time allows"

KEEP IN MIND: these 20 technical courses are already outliers for including ethics in their syllabi at all -- Most, if not all, technical courses approached ethics as a mathematical concept instead of looking at AI Ethics from the social context

Table 4 outlines various ethical topics not given enough attention, or any attention at all, in either technical or non-technical AI courses. The design of AI systems that take into account users with disabilities such as blindness and deafness, was only included in one out of 51 AI and ML syllabi with ethical components. Exposure to the ethical concepts of accessibility is important to ensure that future AI systems are designed and deployed inclusively, not leaving out marginalized groups. While diversity and inclusivity in the AI workforce is given no explicit coverage in the AI ethics curricula, these topics could be included elsewhere in the curricula, such as assembling diverse engineering and design teams, without being explicitly tied to "AI ethics." While military applications of AI are given attention in non-technical courses, it is only considered through the lens of lethal autonomous weapons (LAWS), which is quite limited and does not fully encompass the scale of potential military applications of AI. Of note is the disregard of the environmental consequences of widescale development and deployment of AI. Researchers at the University of Massachusetts Amherst released a report that estimated that the amount of power required to develop and train a neural network architecture involves roughly 626,000 pounds of carbon dioxide, which amounts to nearly five times the emissions of the average American-made automobile across its entire life cycle.[8]

Table 4: Ethical Topic Areas Not Covered

WHAT'S MISSING?	
Accessibility	1 out of 51 courses mentioned accessibility for people with disabilities
Diversity in the AI Workforce	This was not addressed specifically in syllabi / lack of progress to addressing the diversity & inclusion issues within tech companies
Human Element of AI Development	Outside of adjacent discussion on topics such as the future of work, the human element of the development of AI, such as the legions of data labelers/taggers, or issues related to human/machine teaming, are not covered independently

[8] Strubell, E., Ganesh, A., and McCallum, A. (2019). "Energy and Policy Considerations for Deep Learning in NLP." *University of Massachusetts Amherst*. https://arxiv.org/pdf/1906.02243.pdf.

Sustainability	Environmental impact of AI and computing generally is not addressed even though this issue will likely increase in importance
Military AI	Beyond LAWS

Obstacles & Barriers to Embedding AI Ethics in Technical Curricula

The primary obstacle to integrating ethics in technical computer science courses is that instructors without a background in the area might not feel qualified, or comfortable, teaching ethics.[9] Ethics can be difficult to define and evaluate, which may lend towards technical instructors being less inclined to incorporate disciplines in their teachings that are not as concrete or formulaic as mathematics.[10]

Another potential barrier is the misconception that applied ethics education does not translate into actionable practices outside of the classroom. This perception is codified in the Accreditation Board for Engineering and Technology (ABET) Engineering Criteria 2000 (EC2000) accreditation criteria that focuses on ensuring programs provide graduates with the technical and professional skills employers demand.[11] While understanding professional and ethical responsibility became an explicit outcome of university engineering program accreditation, graduates are then faced with entering a technology workforce dominated by generating revenue, user engagement, and increasing value for stockholders.[12] Given that ethics training in technical curricula is limited, and the technology workforce is largely driven by increasing revenue, it comes as no surprise that a North Carolina State University study found that technology companies' codes of ethics do not influence project design decisions of software developers.[13]

[9] Fiesler, C., Garrett, N., and Beard, N. (2020). "What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis." *SIGCSE*; March. https://cmci.colorado.edu/~cafi5706/SIGCSE2020_EthicsSyllabi.pdf.

[10] Polmear, M., Bielefeldt, A. R., Knight, D., Swan, C., Canney, N. E. (2018). "Faculty Perceptions of Challenges to Educating Engineering and Computing Students About Ethics and Societal Impacts." *Paper presented at 2018 ASEE Annual Conference & Exposition*, Salt Lake City, Utah. https://peer.asee.org/faculty-perceptions-of-challenges-to-educating-engineering-and-computing-students-about-ethics-and-societal-impacts.

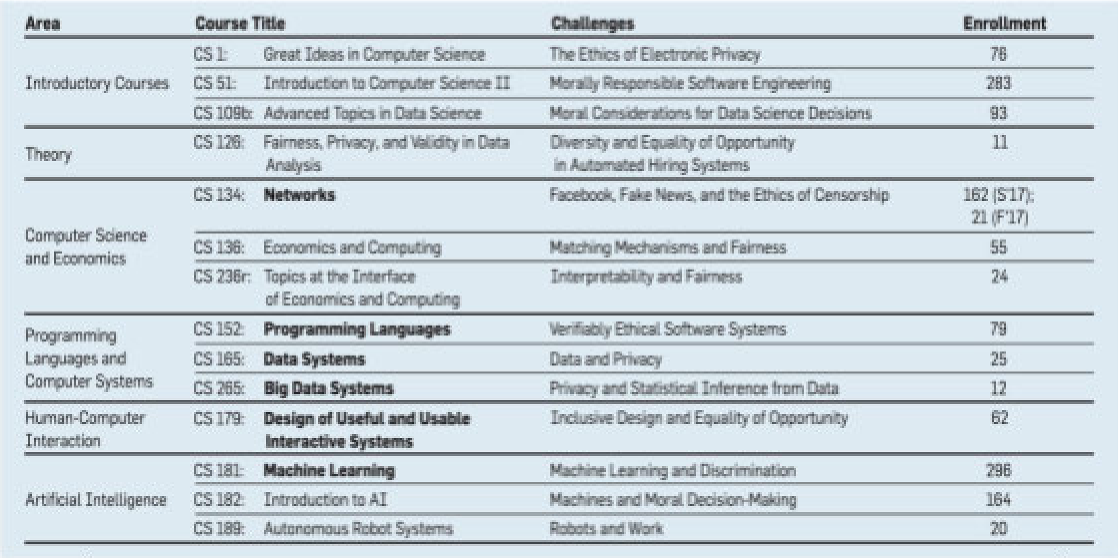
[11] For more information on ABET EC2000, see ABET's website, available at: https://www.abet.org/about-abet/history/. Hess, J. and Fore, G. (2018). "A Systematic Literature Review of US Engineering Ethics Interventions." *Sci Eng Ethics*; 24: pp 551-583. https://link.springer.com/content/pdf/10.1007/s11948-017-9910-6.pdf also acknowledges the role ABET EC2000 plays in the development of pedagogical techniques that shape students' exposure to ethics in technical curricula.

[12] Metcalf, J. Moss, E., and Boyd, D. (2019). "Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics." *Data & Society* originally appeared in *Social Research: An International Quarterly*; 82(2): 449-476. https://datasociety.net/wp-content/uploads/2019/09/Owning-Ethics-PDF-version-2.pdf.

[13] North Carolina State University, (2018). "Code of ethics doesn't influence decisions of software developers." *Science Daily*; October. https://www.sciencedaily.com/releases/2018/10/181009113617.htm. For more information about big technology workforce perceptions of the industry and their employers, see: Baron, J. (2019) "Tech Workers Are Still Willing to Work for Scandal-Ridden Companies," *Forbes*; 4 March.

Embedded EthiCS is a Harvard-based pilot program that aims to address some of these obstacles by integrating ethics throughout the standard CS coursework. The initial idea underpinning Embedded EthiCS began when Dr. Barbara Grosz designed a new course in 2015 titled, “Intelligence Systems: Design and Ethical Challenges.” Co-taught by colleagues in the philosophy department, the course quickly attracted more than 140 students competing for 30 seats in its second year.[14] Four years later, Embedded EthiCS has expanded to a dozen courses in the CS department and will begin extending the model to other disciplines in the near future.[15] Figure 1 below lists the courses, grouping them by CS area, indicating the ethical problems addressed and enrollments.

Figure 1: Embedded EthiCS Courses 2017-2018



Area	Course Title		Challenges	Enrollment
Introductory Courses	CS 1:	Great Ideas in Computer Science	The Ethics of Electronic Privacy	76
	CS 51:	Introduction to Computer Science II	Morally Responsible Software Engineering	283
	CS 109b:	Advanced Topics in Data Science	Moral Considerations for Data Science Decisions	93
Theory	CS 126:	Fairness, Privacy, and Validity in Data Analysis	Diversity and Equality of Opportunity in Automated Hiring Systems	11
Computer Science and Economics	CS 134:	**Networks**	Facebook, Fake News, and the Ethics of Censorship	162 (S'17); 21 (F'17)
	CS 136:	Economics and Computing	Matching Mechanisms and Fairness	55
	CS 236r:	Topics at the Interface of Economics and Computing	Interpretability and Fairness	24
Programming Languages and Computer Systems	CS 152:	**Programming Languages**	Verifiably Ethical Software Systems	79
	CS 165:	**Data Systems**	Data and Privacy	25
	CS 265:	**Big Data Systems**	Privacy and Statistical Inference from Data	12
Human-Computer Interaction	CS 179:	**Design of Useful and Usable Interactive Systems**	Inclusive Design and Equality of Opportunity	62
Artificial Intelligence	CS 181:	**Machine Learning**	Machine Learning and Discrimination	296
	CS 182:	Introduction to AI	Machines and Moral Decision-Making	164
	CS 189:	Autonomous Robot Systems	Robots and Work	20

SOURCE: Grosz, B.J., Grant, D.G., Vredenburgh, K., Behrends, J., Hu, L., Simmons, A., Waldo, J. (2019). “Embedded EthiCS: Integrating Ethics Across CS Education.” *Communications of the ACM*, 62 (8): p. 57. https://dl.acm.org/doi/pdf/10.1145/3330794.

The program employs a distributed pedagogy that makes ethical reasoning an integral component by modifying existing courses rather than requiring wholly new standalone courses. Short ethics modules are distributed throughout CS courses in the core curriculum, and philosophy faculty have partnered with their CS counterparts to develop the Embedded EthiCS curriculum. Importantly, the pioneers of Embedded EthiCS have made the course materials

https://www.forbes.com/sites/jessicabaron/2019/03/04/tech-workers-are-still-willing-to-work-for-scandal-ridden-companies/#47ec36b43c29.

[14] Karoff, P. (2019). “Harvard initiative seen as national model.” *The Harvard Gazette*; 25 January. https://news.harvard.edu/gazette/story/2019/01/harvard-works-to-embed-ethics-in-computer-science-curriculum/.

[15] Hailu, R.A., and Jia, A.K. (2019). “Joint CS and Philosophy Imitative, Embedded EthiCS, Triples in Size to 12 courses.” *The Harvard Crimson*; 22 February. https://www.thecrimson.com/article/2019/2/22/embedded-ethiCS-expands/.

available online in an open-source format[16] so students, faculty members, and other institutions can easily access them to increase the potential for wide-range adoption at any university.

At the Worcester Polytechnic Institute (WPI), The "Great Problems Seminars," which address a range of global sociotechnical problems, such as AI ethics, and the "Grand Challenges" Scholars Program Framework were efforts to increase partnerships between humanities and social science instructors and STEM instructors and produce modules that could be used across departments.[17] Both the Harvard and WPI efforts raise an important point – what impact does the absence of ethics courses and topics have on technical fields and how can other departments responds to fill the gaps?

Obstacles & Barriers to Embedding AI Ethics in Non-Technical Curricula

Just as it is crucial to embed ethics in technical AI curricula, it is also important to integrate technical dimensions of AI in non-technical courses. The intersection and convergence of science, technology, and the humanities have shaped societies throughout history. Digital humanities – the use of digital tools, methods, or approaches to extend the human capacity to explore questions relating to people, cultures, or communities[18] – is gaining steam at various colleges and universities and aims to inject pedagogy, ethics, social context, and creativity across students' educational experience.[19] Increasingly, non-technical disciplines such as political science, philosophy, anthropology, and sociology will need to incorporate data science and AI into the curricula in order to keep pace with the digitization of research, analysis, presentation, and dissemination.[20]

Barriers to integrating AI ethics into non-technical curricula mirror the obstacles to embedding ethics in technical courses. Humanities instructors face similar challenges overcoming perceived, or real, lack of knowledge about data science and AI. However, non-technical

[16] For more information about the Embedded EthiCS module, see: https://embeddedethics.seas.harvard.edu/module.html.

[17] National Academies of Sciences, Engineering, and Medicine (NASEM). (2018). *The Integration of the Humanities and Arts with Sciences, Engineering, and Medicine in Higher Education: Branches from the Same Tree*. National Academies Press: p. 74. https://www.informalscience.org/sites/default/files/24988.pdf.

[18] See the *Report of the Working Group on Data Science and the Digital Humanities* available at: https://www.williams.edu/strategic-planning/files/2020/02/Data-Science-and-Digital-Humanities-report-Technology-and-the-Liberal-Arts.pdf: p. 4.

[19] For more information on digital humanities and how one college is working to integrate technical and non-technical aspects of the curriculum, see the *Report of the Working Group on Data Science and the Digital Humanities* available at: https://www.williams.edu/strategic-planning/files/2020/02/Data-Science-and-Digital-Humanities-report-Technology-and-the-Liberal-Arts.pdf.

[20] *Report of the Working Group on Data Science and the Digital Humanities* available at: https://www.williams.edu/strategic-planning/files/2020/02/Data-Science-and-Digital-Humanities-report-Technology-and-the-Liberal-Arts.pdf:

instructors are likely aware of the need for their students to have training in technical systems as they enter the workforce.

How AI ethics is embedded in both technical curricula and non-technical curricula is critical. Table 5 depicts the pedagogical interventions referenced in J.L. Hess and G. Fore's 2018 "Systematic Literature Review of US Engineering Ethics Interventions"[21], organized first by percentage of mentions in the literature and second by whether the interventions were developed by the instructors or the students.

Table 5: AI Ethics Pedagogy

Course Activities	
Ethical Codes, Rules, and/or Guidelines	85%
Cade Study Exposure	81%%
Ethical Heuristics	46%%
Philosophical Ethics / Theoretical Grounding	42%%
Community Engagement / Service Learning	8%
Inductive Discussion and/or Debate	77%
Individual Written Assignments	54%
Peer Mentoring	12%
Micro-Insertion	8%
Game-Based Pedagogy	8%
Real-World Exposure	8%
Student Involvement in Pedagogy Development	
Case Studies	12%
Ethical Heuristics	12%
Ethical Codes, Rules, or Guidelines	8%

This is important because effective pedagogical strategies identified transfer well between technical and non-technical curricula. For instance, game-based pedagogy could be employed by policy and public affairs instructors in the form of a policy hackathon that requires non-technical students applying machine learning to policy narrative development.[22]

[21] Hess, J. and Fore, G. (2018). "A Systematic Literature Review of US Engineering Ethics Interventions." *Sci Eng Ethics*; 24: pp 551-583. https://link.springer.com/content/pdf/10.1007/s11948-017-9910-6.pdf.

[22] One example of such a policy hackathon was conducted by the Pardee RAND Graduation School in 2019, which asked participants to "tell a useful, interesting, or illuminating story about vendor behaviors and offerings on dark web cryptomarkets using machine learning to develop models to improve the classification of goods to identify vendor patterns." More information on this policy hackathon can be found at: https://www.prgs.edu/news/2019/dark-web-hackathon.html.

A Look into AI Ethics within Pardee RAND Graduate School (PRGS)

After identifying and assessing prior work conducted by Garrett at al. and Saltz et al., we applied their data analysis methodology to the 2019-2020 PRGS curricula in order to determine the current level of AI ethics integration as well as entry points for how PRGS might be able to further embed AI ethics. We first compiled the PRGS syllabi for 32 courses taught between Fall 2019 and Spring 2020, and then we mined each for topics listed in the schedule, reading list, or listed in the course description and conducted content analysis for integration of AI ethical topics. While not all syllabi included all of the components listed above, Garret et al. and Saltz et al. faced the same limitation. Below we will discuss the results of this content analysis of the syllabi.

Table 6: Ethical Components in the 2019-2020 PRGS Curricula

		% of Syllabi
Race (inequality, insecurity, class, etc.)		38%
Gender (inequality, insecurity, etc.)		31%
Bias		25%
Behavioral Economics (nudges, etc.)		25%
Military AI beyond LAWS (cyberwar, etc.)		19%
Philosophy (Kant, Hobbes, Meta-Ethics, Heuristics, individualism v. collectivism, justice, etc.)		19%
Predictive Policing		13%
Climate change		13%
Sexuality & Sex-Selected Behaviors		13%
Tech Ethics		6%
Ethical codes		6%
Accountability		6%
Explainability		6%
Fairness		6%
Privacy (near-term)		6%
Surveillance		6%
Social Media – security, democracy, privacy		6%
Existential risks (far-term)		6%
Well-being		6%
Other	Conducting ethical economic research	6%
	Social determinants of health	13%
	Defining ‘good performance’ in a system	6%
	Protection of human research subjects training	6%
	Logic models	6%
	Abortion	6%

NOTE: Rows highlighted in purple are AI ethics topic areas in PRGS syllabi also identified in the Garrett et al. and Saltz et al. studies. The rows not highlighted are AI ethical topic areas missing from those studies but appear in the 2019-2020 PRGS curricula.

While the ethical components included within the 2019-2020 PRGS curricula largely follow those found in the analyses conducted by Garret et al. and Saltz et al., PRGS does include some of the missing topic areas identified earlier (e.g., military AI beyond LAWS, climate change, etc.). However, none of the topic areas in the PRGS curricula approach the percentage of inclusion in the syllabi found in other studies.[23]

The 2019-2020 PRGS curricula included a total of 32 courses, ranging from microeconomics and math for policy analysis to inequalities in social policy and machine learning (see Table 7).

Table 7: PRGS 2019-2020 Curricula Analysis

Course #	AI Ethical Topics Covered	Ethical Topics Included	Co-Taught	Technical Course	Qualitative Course	Entry-Point
1	Y	Y	Y	N	Y	-
2	Y	Y	N	N	Y	-
3	Y	Y	Y	N	Y	-
4	Y	Y	Y	N	Y	-
5	Y	Y	N	/	/	-
6	Y	Y	Y	/	/	-
7	N	Y	N	N	Y	Y
8	N	Y	Y	N	Y	Y
9	N	Y	N	N	Y	Y
10	N	Y	Y	N	Y	N
11	N	Y	N	N	Y	N
12	N	Y	N	/	/	N
13	N	Y	Y	N	Y	N
14	N	Y	N	/	/	Y
15	N	Y	Y	/	/	Y
16	N	Y	N	N	Y	Y
17	N	N	N	Y	N	N
18	N	N	N	/	/	Y
19	N	N	N	/	/	N
20	N	N	N	N	Y	Y
21	N	N	N	Y	N	N
22	N	N	N	/	/	N
23	N	N	N	Y	N	N
24	N	N	N	Y	N	N
25	N	N	Y	N	Y	Y

[23] One limitation that affects the 2019-2020 PRGS syllabi content analysis that did not appear in the Garret et al. or Saltz et al. studies is that PRGS is a policy and economics graduate program. While Garret at al. and Saltz et al. were able to narrowly tailor their dataset to only include AI/ML courses, the same is not possible for the content analysis of the PRGS curricula. This limitation impedes further comparison between ethical topic area content analyses within the syllabi.

26	N	N	Y	Y	N	Y
27	N	N	N	/	/	N
28	N	N	Y	/	/	N
29	N	N	N	/	/	N
30	N	N	N	/	/	N
31	N	N	N	/	/	N
32	N	N	Y	/	/	Y

NOTE: Courses 1-6 in this tables are the PRGS courses that have integrated AI ethics into its coursework. Courses 7-16 are courses that have integrated ethical components, but not necessarily specific to AI ethics. Y (green) indicates yes; N (red) indicates no; / (grey) indicates mixed method courses; - (white) indicates no entry-point necessary**.** Entry points were determined based on the content analysis of the course syllabi. Entry point was indicated as a no if the course was not structured to allow for either an additional instructor or the material was not applicable to the integration of AI ethics.

Of those 32 courses, approximately 19% integrated AI ethical topics (e.g. predictive policing, race/gender discrimination/bias, fairness, explainability, lethal autonomous weapons, privacy, surveillance, etc.) Of those six courses that integrate AI ethical topics, two courses are standalone technical courses and the remaining four courses are either qualitative or mixed method. While approximately 84% of PRGS courses are either qualitative or employ a mix-method approach to policy problems, 16% of the 2019-2020 PRGS curricula are standalone technical courses.

Approximately 34% of all 2019-2020 PRGS courses are taught by multiple instructors. Of the courses that have embedded AI ethical topics, 67% are co-taught. This creates a variety of entry-points for PRGS leadership to further embed ethical topics throughout the PRGS curricula. For example, standalone qualitative courses are more frequently co-led than either mix-method or standalone technical courses, which suggests that co-led courses are able to more easily integrate AI ethical topics into coursework. If mix-method and technical PRGS course instructors were encouraged to reach out to RAND colleagues with background in ethical research, and co-develop and teach these courses, then PRGS may be able to avoid the barriers and obstacles mentioned above – such as including AI ethics "if time allows" and concerns expressed by technical instructors that they are not comfortable teaching topics outside of their field of expertise.

Next Steps

This analysis will inform the development of a curriculum offered through the Pardee RAND Graduate School intended to facilitate discussion among both RAND researchers and Pardee RAND students about the ethics of AI. The goal of providing the curriculum for RAND researchers, in addition to Pardee RAND students, is to promote the integration of ethical analysis into RAND's research and policy recommendations for government and other sponsors related to AI.

The annotated syllabus is designed to address and overcome both the barriers to the integration of ethics in technical courses as well as the integration of AI in non-technical disciplines discussed in this analysis. The syllabus was developed with the intent that it could be used both within and outside of RAND, and scholars at other institutions could use it to identify topics and readings that may not have otherwise considered including in their curricula.

Beyond the annotated syllabus, and in support of integrating AI ethics into PRGS activities, the research team is launching an ethics hackathon in which PRGS students will work in teams to address the ethical implications associated with a COVID-19 tracking platform developed by the non-profit COVID-Alliance. The hackathon will launch on 20 July 2020 and run for three weeks. The Pandemic Management Platform (PMP) provides live geolocation data, hospital capacity data, infection data, and other information to U.S. decision-makers to enable effective public health interventions. PRGS students will integrate their technical knowledge and skills with non-technical ethical considerations such as – what are the potential ethical harms of the platform and how can they be mitigated through design features, visualizations, and use-restrictions?

Considerations

This memo should be considered in light of several considerations and limitations. For example, we have not replicated the qualitative studies referenced herein and have incorporated findings from other works on this topic. The percentages in Tables 1-4 should not be taken as hardline quantitative representation of levels of inclusion or interest. Rather, the percentages are more descriptive of the overall space of ethical topics included in AI Ethics syllabi. Since Table 5 was derived from a systematic literature review, those percentages are more indicative of actual levels of inclusion of various pedagogical interventions.

Annex 2

An Annotated Syllabus for AI Ethics

Version: September 30, 2020

As artificial intelligence (AI) becomes increasingly prominent across a range of consumer applications and is integrated into high stakes social and political institutions, there has been increased attention to the ethical considerations related to its development and use. This document provides an annotated syllabus for a curriculum on AI ethics by identifying key AI ethics topics and readings.

The objectives of this document include:

- Building a modular curriculum on issues in AI ethics that provide a foundation for coursework and that enables independent self-study.
- Presenting a variety of topics and a breadth of readings that provide an overview of key issues and additional references.
- Initiating reflection by raising important questions and putting readings in dialogue with one another.

Readers may be familiar with the oft-quoted line from *Jurassic Park*, when a character admonishes the creator of the park by stating, *"Your scientists were so preoccupied with whether or not they could, they didn't stop to think if they should."* This key distinction – between the "could" and "should" in technological progress – underlies many of the ethical issues raised in the design, development, and implementation of AI. That distinction led to several overarching questions that guided the development of this syllabus.

1. How should we balance innovation and technological exploration with ethical considerations and related values such as responsibility, equity, and well-being? What are the ethical risks and benefits of new technologies and how can potential harms be assessed and addressed?
2. Who is responsible for considering the ethics of technological developments and their application, and how should this be done? What are the mechanisms for ensuring critical ethical engagement during the development, testing, and application of new technologies? What should be done when harms are not recognized until after deployment?
3. How do the cultural, political, demographic, and economic contexts in which these technologies are developed, designed, and deployed shape the benefits, risks, and harms? How do they prompt the ethical, legal, and social responses to these technologies?

The main topics in AI ethics were chosen due to the importance of the issues and the maturity and range of research related to them. Though the syllabus is not intended to cover every aspect of AI ethics (or closely related fields such as Internet or cyber ethics), it does provide a broad introduction to many of the major topics discussed in the literature. The syllabus presents only a temporal snapshot of a rapidly evolving field, so it will need to be regularly updated to incorporate new issues and perspectives.

Additional resources related to AI ethics can also be found through the following organizations, each of whom has done important work to advance the field:

- The Algorithmic Justice League
- The AI Now Institute
- Google Responsible AI
- Markkula Center for Applied Ethics
- Microsoft Responsible AI
- Data for Black Lives
- Stanford Human-Centered AI
- Center for Humane Technology
- ACM FAccT

This is just a sampling of key organizations and research, and there are many other resources available, some of which are further described below.

Topics/Modules

Ethics of Technology Overviews

One often-heard claim is that technology is value neutral, existing simply as an inert artifact that does not inherently embody any normativity or ethical salience on its own. On this view, technology ethics does not fundamentally concern the technology itself, but the ways humans use technology in practice. Rather than slow-down or critique the development of technology, ethical reflection on technology should focus on how humans use it. This view is regularly articulated in the context of AI, where some argue that AI is just math—a form of advanced mathematical statistics—and should not be subject to ethical critique in and of itself. Similarly, the data that provides the key fuel for improving AI is objective and unbiased, and might not itself be the subject of ethical critique. However, an alternative view is that AI artifacts and their underlying data sources themselves reflect the values of its creators and the social and the broader cultural institutions in which they are found, developed, and implemented. On this latter view, there are important questions about the ethics that are infused within the technology and data, and undergirds work that seeks to ensure that the development of AI is done with ethical constraints in mind. An additional set of issues concerns the perspectives from which one approaches the ethics of AI. Much research is from the standpoint of "western" philosophical approaches and values that does not account for the full global lens under which ethical reflection proceeds.

Key questions:

1. Is AI "value-neutral" or does it inherently embody human values and perspectives?
2. Can ethical norms be directly programmed into AI?

3. Are there differences between so-called Western and other approaches to the ethics of AI?

- Shannon Vallor et. al "Overview of Ethics in Tech Practice" https://www.scu.edu/ethics-in-technology-practice/overview-of-ethics-in-tech-practice/
    - o A short reference guide that offers some key distinctions (eg between ethics and compliance) and suggests a set of resources and tools that can foster reflection and help to integrate ethical thinking into technology practice.

- Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, Rashida Richardson, Jason Schultz, Sarah Myers West, and Meredith Whittaker. AI Now 2019 Report. New York: AI Now Institute, 2019, https://ainowinstitute.org/AI_Now_2019_Report.html.
    - o The AI Now Institute's annual report explores the harms and accountability gaps of AI tools such as surveillance technology, including exacerbating inequity and power asymmetries. It describes efforts by a range of coalitions to draw attention to these risks, and offers a set of recommendations to improve the accountability, transparency, and legitimacy of AI technology.

- Langdon Winner "The Whale and the Reactor: A Search for Limits in an Age of High Technology" The University of Chicago Press, 1996
    - o First published in 1988, Winner provides a rich historical exploration of the ways in which decisions about technical development are themselves "political" in nature, involving considerations related to status, power, and justice. These considerations are embedded into the technological artifacts that surround us. Technology is presented not just as a set of tools for human use, but as enabling and constraining factors that open or foreclose human possibilities for action. In this way, technology is fundamentally intertwined with human agency and meaning. The work provides a subtle philosophical analysis that upholds the need to continually interrogate and question modern technological developments.

- Montreal AI Ethics Institute: **"**The State of AI Ethics" June 2020 **https://arxiv.org/pdf/2006.14662.pdf**
    - o A "pulse check" on the discourse surrounding AI ethics, including issues such as misinformation, privacy, and the future of work. The report presents different perspectives on each of these issues along with extensive resources to go deeper.

- Vaibhav Garg, L. Jean Camp "Gandhigiri in cyberspace: a novel approach to information ethics" ACM SIGCAS Computers and Society, August 2012 **https://dl.acm.org/doi/abs/10.1145/2422512.2422514**

    - Develops a Gandhi-inspired approach to information ethics and contrasts this approach with the traditional Western approaches that tend to get greater attention in the literature.

- Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*, 2019.
    - Written by a well-regarded computer scientist who has contributed significantly to AI, this general audience work underscores the risks and dangers to humanity of AI that is not aligned to human values. Russell criticizes the traditional approach to AI that focuses on building systems that can optimize well-defined and fixed human goals. Instead, Russell argues AI should be designed to be deferential to uncertain and shifting human values. Rather than be given a fixed goal, the objectives of AI systems should be to maximize human preferences, while proceeding under uncertainty as to what those preferences ultimately are.

Diversity in AI

Discussions of diversity in AI often engage with two themes: how the implementation of AI causes harm based on discriminatory bias embedded in algorithms or in AI design and how a lack of diversity within tech companies is problematic, in part because it leads to that embedded bias. The selected readings review both of these themes in-depth – documenting the impacts of discriminatory practices, whether intentional or not, and offering solutions on how to address the negative impacts of bias in AI and improve diversity.

Questions

1. How can stakeholders (e.g., designers, developers, policymakers, etc.) address discriminatory bias in AI?
2. Why should companies and policy entities pursue hiring and retention practices that build and maintain a diverse workforce – especially in tech?
3. How might marginalized communities advocate for “a seat at the table”? Is having a seat enough to mitigate the risks and harms associated with AI?

- Gebru, T. (2020). Race and Gender. *The Oxford Handbook on Ethics of AI*, edited by M. D. Dubber, F. Pasquale, and S. Das. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780190067397.013.16
    - Gebru examines the rapid permeation of AI into society, how it can and has caused harms, and discriminatory practices in tech in her entry on “Race and Gender” in *The Oxford Handbook on Ethics of AI*. The chapter notes that more thorough investigations of the sociopolitical issues that lead to harms rather than advantages of AI are needed, and that they should take a holistic and multifaceted approach. Gebru lays out specific examples of how AI can cause harm in order to support the recommendations for developing ethical AI.

- Yeung, D. (2018). “When AI Misjudgment Is Not an Accident.” *Scientific American*, October 19. Retrieved from:

https://blogs.scientificamerican.com/observations/when-ai-misjudgment-is-not-an-accident/

  - Yeung starts with noting that bias in artificial intelligence is often thought of as "unconscious", with a focus on "algorithms that unintentionally cause disproportionate harm to entire swaths of society". However, this piece engages with the ways in which actors may *intentionally* introduce bias into AI systems, the reasons why someone might pursue discriminatory practices, and the impacts that bias could have. Yeung argues that bias is thus a systemic challenge, which requires holistic solutions such as organizational training to identify and mitigate it and legislative oversight.

- West, S.M., Whittaker, M. and Crawford, K. (2019). Discriminating Systems: Gender, Race and Power in AI. *AI Now Institute*. Retrieved from: https://ainowinstitute.org/discriminatingsystems.html.
  - West et al. engage with the diversity crisis in the AI sector, specifically across gender and race. Their report demonstrates how current efforts to address issues of diversity in AI are insufficient, and that of AI systems for the classification, detection, and prediction of race and gender are in urgent need of re-evaluation. Based on the research they present, the authors make several recommendations for improving workplace diversity and for addressing bias and discrimination in AI systems.

- Buolamwini, Joy. (2017). How I'm fighting bias in algorithms [Video file]. Retrieved from: https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms.
  - Buolamwini's TED Talk provides a quick introduction to the types of exclusion and discriminatory practices that can result from algorithmic bias. She also introduces "incoding" (inclusive coding), founded on three principles: who codes matters, how we code matters, and why we code matters. Her details the actions she takes in her efforts to identify bias, curate inclusively, and develop technologies conscientiously by considering their social impacts.

- Houser, K. A. (2019). "Can AI solve the diversity problem in the tech industry? Mitigating noise and bias in employment decision-making." *Stanford Technology Law Review, 22(2),* 290-354. Retrieved from: https://law.stanford.edu/publications/can-ai-solve-the-diversity-problem-in-the-tech-industry/
  - Houser makes a case for "responsible AI" and its potential to mitigate the problems caused by unconscious biases in human decisionmaking in order to increase the hiring, promotion, and retention of women in the tech industry. She reviews new solutions for how AI could be incorporated into decisions that impact employment, with particular attention to the legal dimensions of this type of AI implementation.

Fairness, Equity, and Bias in AI

Despite the contention of some that AI is just an inert technical artifact, there is increased recognition of the multitude of ways that AI might perpetuate unfairness and inequity. These concerns are worsened by the inclusion of AI in high stakes decisions, for instance in housing, criminal justice, health care, or employment. The source of AI unfairness might be traced to the homogeneity and lack of imagination of AI developers, issues related to biased training data that does not sufficiently represent all relevant demographic groups, the ways that AI systems are integrated in existing social structures that are themselves unfair, or even other potential sources. Researchers have approached this topic from a variety of perspectives, some of whom have shown mathematically the ways that AI systems are unfair and articulated different types of fairness criteria that AI should aspire to satisfy. Others have looked at specific social institutions—such as employment—and shown how AI might foster unfairness or discrimination within that institution. Still others have taken a broader socio-technical perspective to underscore ways that AI is part of the systemic challenges of inequity.

Key questions:

1. How should equity and fairness be conceptualized in the context of AI development and deployment?
2. How does AI systemically exacerbate or address forms of inequity and unfairness?
3. How, if at all, can fairness be described formally/mathematically, or otherwise used to assess or constrain AI systems?

- Narayanan, Arvind. “21 Definitions of Fairness and their Politics” 2018 Conference on Fairness, Accountability, and Transparency (FAT*) conference. https://www.youtube.com/watch?v=wqamrPkF5kk
    - A tutorial from the 2018 FAT* conference that lays out 21 separate formal definitions of fairness, some of the underlying intuitions and principles underlying these conceptions, and the impossibility of achieving them simultaneously. The talk underscores why formal or mathematical approaches to fairness will likely be contested.
- Hao, Karen and Jonathan Stray “Courtroom Algorithm Game”, MIT Tech Review, October 17, 2019 https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/
    - A user-friendly guide that depicts competing conceptions of algorithmic fairness in the context of machine learning tools to conduct criminal risk assessment. This tool underscores the need for a richer conception of fairness that is appropriate for the specific social and institutional context.

- Benjamin, Ruha. *Race After Technology.* Polity June, 2019
    - Benjamin takes a systematic and historically-informed approach to underscore how technology that might appear to be race-neutral or even equitable can foster White supremacy and worsen social inequity. She offers insightful critiques about a range

of AI related technologies—from search to risk assessment algorithms—describing them as part of a "new Jim Code."

- Crawford, Kate "The Trouble with Bias" NIPS 2017 Keynote https://www.youtube.com/watch?v=fMym_BKWQzk
  - Crawford's presentation presents an overview of ways that algorithms have exacerbated discrimination and unfairness, and distinguishes between different type of harms associated with AI systems.

- Birhane, Abeba and Fred Cummins "Algorithmic Injustices Towards a Relational Ethics" Black in AI Workshop, NeurIPS2019
  https://arxiv.org/abs/1912.07376
  - The paper reviews existing research showing the ways that algorithms integrated within social institutions foster racial and other forms of injustice and harm marginalized populations. It argues that algorithmic fairness cannot be seen through a purely technical lens, and develops a 'relational' view of fairness that focuses on vulnerable populations.

- Barocas, Solon and Andrew Selbst "Big Data's Disparate Impact" California Law Review, Vol. 104, No. 3, June 2016
  http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf
  - Considers the multiple ways that machine learning algorithms inherit the biases within training data, with a specific focus on algorithms in the employment context. The paper offers a detailed legal review of American discrimination law and the challenges of applying existing disparate treatment and disparate impact criteria to the use of algoritghms.

- Osoba, Osonde et al "Algorithmic Equity: A Framework for Social Applications" RAND Corporation, 2019
  https://www.rand.org/pubs/research_reports/RR2708.html
  - An overview of the problems associated with algorithmic fairness, with a specific focus on the use of algorithms in employment, criminal justice, and insurance contexts. The report makes the case for a broader socio-technical systems approach to identifying and mitigating inequity in the use of algorithms.

- Noble, Safiya Umoja *Algorithms of Oppression*: *How Search Engines Reinforce Racism,* NYU Press, 2018
  - Noble's book shows the ways that bias and unfairness spreads across the Internet. Far from being a purely neutral tool, prominent search engines infuse misogynist and racist values within them. These tools

Privacy in AI

Privacy, sometimes conceptualized as the right to be left alone, is a longstanding and widely resonant value of special importance to vulnerable and marginalized populations. Digital technology and AI have heightened the risks to privacy through the widespread collection of personal sensitive information and powerful algorithms that enable a detailed and potentially intrusive picture of personal lives. However, calculating the ethical risks baked into the use of AI and in particular the role of companies in fostering 'surveillance capitalism' is a challenging question, especially when consumers "consent" to using products that track personal information. Scoping this issue for classroom discussion can be difficult as there are a range of technical approaches to protecting privacy and a robust literature on the legal frameworks in place that apply to data collectors. This section thus focuses more narrowly on the type of risks presented by AI and related technologies, and provides a foundation for some of the approaches to protect personal privacy.

Key questions:

1. What rights do individuals have in regard to their data?
2. What limits should governments place on the use of data, and how should those laws be enforced?
3. What characteristics define various countries' approaches to privacy?
4. How are the benefits and costs of applications driven by big data distributed? What factors drive these tradeoffs?

- Schneier, B. (2015). *Data and Goliath: The hidden battles to collect your data and control your world*. WW Norton & Company.
    - Schneier's book is a widely cited depiction of 'ubiquitous mass surveillance' in the United States and the status of privacy in a world of big data. The first of the book's three parts (which is most highly regarded) focuses on the mechanics of how data is processed by private firms and accessed by government. The subsequent sections describe the societal costs of these programs and policy interventions in response. While alarmist at times, Schneier's expertise serves as a helpful primer in understanding surveillance.

- Jones, M. L., & Kaminski, M. E. (2020). An American's Guide to the GDPR. Denver Law Review, 98(1). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3620198
    - Jones and Kaminski aim to clarify how to understand the GDPR starting from an American legal context. Programs with an emphasis on policy and law will benefit from the overview of data protection and privacy legal regimes and the contrast between the U.S. and E.U. Legal classes might also use the subsequent sections outlining how to interpret and analyze actual GDPR text.

- Gruschka, N., Mavroeidis, V., Vishi, K., & Jensen, M. (2018, December). Privacy issues and data protection in big data: a case study analysis under GDPR. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 5027-5033). IEEE. https://arxiv.org/pdf/1811.08531.pdf

    - o Gruschka et al. outlines the technical approaches to complying with privacy standards in a brief and approachable manner. The article also uses two case studies of Norwegian data sets to demonstrate the inherent tradeoffs involved in ensuring privacy, a helpful read for debating the more specific elements of privacy compliance

- Baik, J. S. (2020). Data privacy against innovation or against discrimination?: The case of the California Consumer Privacy Act (CCPA). Telematics and Informatics, 52. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3624850
    - o More analysis of the CCPA will certainly be published later into its implementation, but Baik's analysis of public comments demonstrates varying narratives and shifts in language between advocates and dissidents of privacy law.

- Zuboff, S. (2019). *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. New York: PublicAffairs.
    - o Zuboff reframes mass surveillance as not merely an abuse of big data practices but rather an entire economic and social system, a 'surveillance-based economic order'. For classes interested in the topic but unable to commit to the 700-page length, journal articles by Zuboff or interviews on her book could provide the core concepts in a more concise manner.

Facial Recognition and Biometric Surveillance

Building on advances in computer vision models and the availability of large training sets, facial recognition technologies (FRTs) have been increasingly integrated into a range of applications, including law enforcement, traveler verification, educational settings, and consumer applications. FRTs are now being used to identify persons, track them over time, determine demographic characteristics (such as age, gender or race), or evaluate affect such as the type of emotion expressed by their face. As organizations have expanded FRT use, criticisms have also surfaced. Some of these critiques have focused on demographic disparities related to the accuracy of these systems, and the invalidity and likely misuse of techniques such as affect recognition. Others have focused on risks to privacy related to the collection and analysis of biometric information, and the chilling effects on civil liberties associated with undermining anonymity in public. In addition, researchers have noted that FRTs are sometimes deployed without sufficient transparency and without proper notification and consent of users. Many critiques have also emphasized that the harms associated with FRTs will have a disparate impact on vulnerable and marginalized persons and will exacerbate racial and social inequity.

Key questions:

1. How is facial recognition used by public and private institutions?
2. What are some of the key risks regarding facial recognition and how can they be mitigated?

- Buolamwini, Joy; Gebru, Timnit, et. al. *Gender Shades Project* http://gendershades.org/index.html

  - A seminal tool and research paper that shows the bias associated with facial recognition technology, and in particular that shows that darker-skinned women are more likely to be mischaracterized by several commercially available facial recognition systems. The research underscores the need for accountability and transparency in the use of these systems.

- Axon Ethics Board “First Report of the Axon AI and Policing Technology Ethics Board” June 2019 https://static1.squarespace.com/static/58a33e881b631bc60d4f8b31/t/5d9df18e9b1895351ceea85f/1570632083376/Axon_Ethics_Report_vfinal-English.pdf
  - A report written by the ethics board of Axon, a major provider of technology to law enforcement, that argues that face recognition technology is not sufficiently reliable for law enforcement use and there is a need for transparent accountability mechanisms.

- Garvie, Clare, A. Bedoya, and J. Frankle. “The Perpetual line-up: Unregulated Police Face Recognition in America” Georgetown University Law School, Washington, DC, 10 2018 https://www.perpetuallineup.org
  - A report laying out the widespread and unregulated use of facial recognition by police across the United States. One of the many findings is that 1 out of every 2 Americans is in a law enforcement face recognition database (derived in many cases from DMV photos), often without transparency or knowledge of the person. The report describes the widespread risks of these practices to privacy and civil rights, and calls on Congress to take legislative action.

- Galligan, Claire, Hannah Rosenfeld, Molly Kelinman, Shobita Parthasarathy “Cameras in the Classroom: Facial Recognition Technology in Schools”
  - A report focused on the use of FRTs in American schools that recommends that the use of FRTs in schools should be banned. The report argues FRTs exacerbate racism, normalize surveillance, commodify private data, and institutionalize inaccuracy.

Policy and Governance of AI

Due to the pace of technological development, stakeholders struggle to create appropriate and effective governance structures and oversight mechanisms to identify, assess, and mitigate risks and harms that may result from new technologies. Those risks and harms are well-documented in the literature and are generally agreed upon. Yet, there are a range of recommendations on how to address the risks and harms, which differ on their framework for defining the issues, their recommendations for how stakeholders should coordinate, and their understanding of the potential results of mitigation efforts. The selected readings present several examples of how to conceptualize the risks associated with technology, and possible solutions for mitigating those risks.

Key questions

1. What are the challenges to building effective governance mechanisms for technology?
2. Identify the different frameworks presented in the readings for providing oversight and regulation of technologies. What are their similarities? Differences?

- Marchant, G., & Wallach, W. (2015). "Coordinating technology governance." *Issues in Science and Technology*, 31(4): 43-50. https://issues.org/coordinating-technology-governance/
  - Marchant and Wallach's article is not specific to policymaking for AI; rather, it engages with questions of governance for emerging technologies in general. This broad approach to considering how to regulate new technologies and their associated risks is important for addressing coordination problems in policymaking. As the authors note, "no single entity is capable of fully governing any of these multifaceted and rapidly developing fields and the innovative tools and techniques they produce." Marchant and Wallach examine what happens when a diverse and numerous group of actors are involved in emerging tech governance – namely that there are inconsistent recommendations, duplication of efforts, and general confusion. As a result, they propose and detail governance coordination committees, and the ways in which these types of governance structures will play a moderating and convening role.

- Citron, D., & Pasquale, F. (2014). "The scored society: Due process for automated predictions." *Washington Law Review,* 89(1): 1-34. https://heinonline.org/HOL/P?h=hein.journals/washlr89&i=8
  - Citron and Pasquale's piece starts with the point that "procedural regularity is essential for those stigmatized by AI scoring systems", which are pervasive and consequential yet opaque and lacking oversight. They note the human bias that becomes embedded in these systems during development, and lay out their argument for why algorithmic scoring should not proceed without expert oversight. They provide recommendations for how to assess the risks associated with these scoring systems and to design and implement regulatory oversight.

- Lepri, B., Oliver, N., Letouzé, E. *et al.* (2018). "Fair, Transparent, and Accountable Algorithmic Decision-making Processes." *Philosophy & Technology,* 31: 611–627. https://doi.org/10.1007/s13347-017-0279-x
  - Similar to Citron and Pasquale, the authors of "Fair, Transparent, and Accountable Algorithmic Decision-making Processes" discuss available solutions to enhance fairness,

  accountability, and transparency in algorithmic decisionmaking. Lepri, Oliver, and Letouzé's focus, however, also includes "technical" solutions, i.e., efforts among technologists to provide transparency and address discriminatory bias in the design and use of algorithmic decisionmaking. While policy solutions are needed as well, the authors argue that governance mechanisms have not and cannot keep pace with technological developments. The article also discusses Open Algorithms (OPAL) – a project that seeks to “enable the design, implementation and monitoring of development policies and programs, accountability of government actions, and citizen engagement while leveraging the availability of large-scale human behavioral data in a privacy-preserving and predictable manner.”

- Rahwan, I. (2018). “Society-in-the-loop: programming the algorithmic social contract.” *Ethics & Information Technology* 20: 5-14. https://doi.org/10.1007/s10676-017-9430-8
  - Rahwan builds on the earlier literature on the need for AI governance by proposing a conceptual framework for the regulation of AI and algorithmic systems. More specifically, he argues that society needs to develop a new social contract – a pact between various stakeholders to be mediated by machines. This new social contract will require the adaptation of the human-in-the loop (HITL) concept from modeling, simulations, and interactive machine learning. In this paper, Rahwan details his “society-in-the-loop” (SITL) proposal, which would combine the HITL control paradigm with mechanisms for negotiating the values of various stakeholders affected by AI systems, in addition to monitoring compliance with the pact.

- Zarsky, T. (2016). “The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making.” *Science, Technology, & Human Values,* 41(1): 118–132. https://doi.org/10.1177/0162243915605575
  - Like the other readings included in this section, this article engages with the ethical issues of algorithmic decisionmaking. Where it differs from the others is in the analytical framework offered to categorize and understand those issues. Zarksy narrows the discussion down to two key dimensions: the specific and novel problems that algorithmic decisionmaking processes generate and the attributes that exacerbate these problems.

- Jones, M. (2015). The ironies of automation law: Tying policy knots with fair automation practices principles. *Vanderbilt Journal of Entertainment & Technology Law,* 18(1): 77-134. https://heinonline.org/HOL/P?h=hein.journals/vanep18&i=88
  - Meg Leta Jones uses case studies to analyze the various legal approaches to automation taken by legislative, administrative, judicial, state, and international bodies. Her approach is different from the problem-centric frameworks, though she does borrow from each of them. Instead, Jones steps back to take a broader view of a range of technologies – new and old – that fall under “automation”. She finds that despite the intent that automation regulation should protect and promote human values, it results in less protection of human values.

Military Applications of AI

Militaries worldwide have studied closely the rapid advances in AI technology by researchers and in the private sector. Several of these developments, such as the use of AI in game-playing, improvements to logistics, supply chains, and enterprise management, and the advancement of autonomous systems and advanced robotics, might be particularly well-suited to advance military objectives. The U.S., China, Russia, and other major global powers have committed significant investment to integrate AI into military applications, including in command and control, targeting and weapons, and business processes. As a result of these developments, advocates have raised alarm that there is a global AI arms race where countries have rushed to integrate AI into military applications without sufficient attention to ensuring the safety and security of these systems. They have noted that the use of AI technologies in military applications raise significant risks, especially when these systems have lethal payloads or are integrated into lethal targeting processes. Further, these advocates have observed that there is not sufficient technical testing and evaluation or policy/regulatory structures in place to ensure that these systems can be used reliably and in accordance with legal obligations, such as the Law of Armed Conflict. The readings in this section provide an overview of the key questions and issues related to military applications of AI.

(1) How are militaries integrating AI and what are some of the key risks?
(2) How can the international community regulate the use of military AI, including autonomous weapons?

- Human Rights Watch “Making the Case: The Dangers of Killer Robots and the Need for a Preemptive Ban” December 2016. https://www.hrw.org/report/2016/12/09/making-case/dangers-killer-robots-and-need-preemptive-ban
  - Argues that autonomous weapons are ethically and legally prohibited, and there is a need for a new global treaty that bans so-called killer robots.

- Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War.* W.W. Norton & Company, April 2018.
  - A sweeping overview of the attempts by militaries worldwide to integrate AI into into weapon systems, that includes detailed analysis of the military import of different AI technology. Also includes an overview of the legal and ethical issues that have been raised.

- Roff, Heather M. and Moyes, Richard. “Meaningful Human Control, Artificial Intelligence and Autonomous Weapons.” Briefing paper prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Con- vention on Certain Conventional Weapons, April 2016. http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf
  - An important early articulation of the standard for “Meaningful Human Control” which has become a key concept in international discussions related to regulating autonomous weapons. The paper lays out a framework for evaluating human control, which includes technical and other requirements embedded across the life-cycle of an AI system and at various levels of warfare.

- Crootof, Rebecca “War Torts: Accountability for Autonomous Weapons”, University of Pennsylvania Law Review, September 2016 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2657680
    - An in-depth legal exploration of the challenge related to human accountability of autonomous weapons. Some advocates against autonomous weapons have argued that there is an accountability gap whereby no person can be held legally or morally accountable for a weapons system that selects and engages targets on its own. This piece describes the need for a broader legal regime of state responsibility to address this problem.

Disinformation and Content Moderation

While new technologies can bring social benefits, they can also introduce significant harms. The ability to realistically alter images, video, and audio, for example, presents new mechanisms that could exploit, intimidate, and sabotage others. The readings selected engage with both the benefits and the potential risks and harms of technologies that facilitate the spread of disinformation. In addition to documenting the associated issues, the readings present and evaluate approaches to moderating content in an effort to mitigate the risks and harms.

Key questions

1. Should platforms be allowed to enjoy “power without responsibility”? Explain your position.
2. If we cannot control the use of the technology once it has been released into the world, what ethical questions should developers consider as they create?
3. How might legal solutions and public policy shape cultural responses to disruptive technologies? Who (i.e., which stakeholders) should be involved in this type of legal or policy development?

Readings

- Kavanagh, J., & Rich, M. D. (2018). *Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life*. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_reports/RR2314.html
    - This RAND Report defines and examines “truth decay” – a term for the trends associated with a shift away from reliance on objective facts in political debates and policy decisions. The authors define the term as a set of four, interrelated trends: 1) increasing disagreement about facts and analytical interpretations of facts and data, 2) a blurring of the line between opinion and fact, 3) the increasing relative volume, and resulting influence, of opinion and personal experience over fact, and 4) declining trust in formerly respected sources of factual information. The objective of the report is to provide a foundation for policymakers, researchers, educators, journalists, and others to analyze the concepts and relationships that might be contributing to truth decay. In addition to defining the concept in this context, the authors examine each of the four trends, their possible causes and consequences, and whether truth decay is a

new phenomenon. The report also offers a four research streams to continue to investigate truth decay. The report provides an in-depth understanding of truth decay, which is critical to understanding the role of technology in information dissemination and the spread of disinformation.

- Heaven, W.D. (2020). OpenAI's new language generator GPT-3 is shockingly good—and completely mindless. *MIT Technology Review*, July 20. https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/
    - Heaven reports on the release of GPT-3, which he calls "the most powerful language model ever." GPT-3 is a big leap forward in development compared to its predecessor (GPT-2), which was already able to produce streams of text in a range of different styles that were convincing as authentic human works. This report provides details on the type of content GPT-3 can produce, and the potential benefits and risks of this technological advancement.

- Chesney, B., & Citron, D. (2019). Deep fakes: looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753-1820. https://dx.doi.org/10.2139/ssrn.3213954
    - Chesney and Citron assess the causes and consequences of disruptive technological change, then engage with the tools – existing and potential – that may be used to mitigate the effects of these technologies. They focus on "deep fakes", which they define as "the full range of hyper-realistic digital falsification of images, video, and audio." And they consider the use of deep fakes in the context of other technologies, like social media platforms, that can create amplifying effects. Finally, Chesney and Citron review technological and legal solutions to regulating the ill effects of deep fakes, although they do not identify any "silver bullet" approaches to reducing risk and harm.

- Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2020). Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes? *The International Journal of Press/Politics.* https://doi.org/10.1177/1940161220944364
    - Dobber et al. engage with the perception that deepfakes are "a powerful form of disinformation", with specific attention as to whether deepfakes impact political attitudes. The authors hypothesized that microtargeting techniques can amplify the effects of deepfakes, by enabling malicious political actors to tailor deepfakes to susceptibilities of the receiver. To test their hypothesis, they created political deepfake and tested its effects on political attitudes in an online experiment. They found that deepfakes generally can have an impact on attitudes toward the targeted politician, though more negative attitudes do not carry over to that politician's party. toward the politician's party remain similar to the control condition. Further examination of "a microtargeted group", however, shows that attitudes toward the depicted politician and their party are significantly lower, suggesting that microtargeting techniques can amplify the effects of a deepfake.

- Pew Research Center (2017). “The Future of Truth and Misinformation Online.” October 19. https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/
    - In summer 2017, Pew Research Center and Elon University’s Imagining the Internet Center conducted a large canvassing of technologists, scholars, practitioners, strategic thinkers and others, asking them to react to the rise of “fake news” and doctored narratives. This report presents the results of the survey.

- Wagner, T. L., & Blewer, A. (2019). “The Word Real Is No Longer Real”: Deepfakes, Gender, and the Challenges of AI-Altered Video, *Open Information Science*, 3(1), 32-46. https://doi.org/10.1515/opis-2019-0003
    - Wagner and Blewer explore how the emergence and distribution of deepfakes could contribute to a potential future in which gendered disparities within visual information production continue to be enforced. The authors reject, however, that a future where deepfakes are “normal” is inevitable, arguing instead that support for feminist-oriented approaches to artificial intelligence-building and a commitment to critical approaches regarding historical biases in media production can deter the distribution of potentially violent, exploitative, and sexist deepfakes. Wagner and Blewer maintain that this approach necessitates a deeper exploration of the role that ethics plays in visual information as a field more generally.

- Yadlin-Segal, A., & Oppenheim, Y. (2020). Whose dystopia is it anyway? Deepfakes and social media regulation. *Convergence*. https://doi.org/10.1177/1354856520923963
    - Yadlin-Segal and Oppenheim perform a comprehensive global review of journalistic discussions on deepfake applications to understand the narratives that have been constructed through media coverage, the regulatory actions associated with these narratives, and the functions that such narratives might serve in global sociopolitical contexts. They find that journalists frame deepfakes as a destabilizing platform that undermines a shared sense of social and political reality, enables the abuse andharassment of women online, and blurs the line between reality and fiction. More specifically, the media’s examination of deepfakes is linked to discussions of dis/misinformation, manipulation, exploitation, and polarization. The authors provide broader practical and theoretical insights about AI content regulation and ethics, accountability, and responsibility in digital culture in light of their findings.

- Roberts, Sarah. *Behind the Screen,* June 2019
    - An important ethnographic study of the humans involved in social media content moderation who are responsible for evaluating user-generated content and making decisions about compliance with company policies. This book is a

necessary look at the many people that sit behind our technology and the toll their role takes on their emotional well-being.

Environmental Implications of AI

The literature presents two ways in which to consider the environmental implications of AI: 1) the potential benefits of its use for achieving sustainability goals and mitigating environmental degradation and 2) the negative environmental impacts of AI's resource use, waste production, and resulting pollution. Like many emerging technologies, these benefits and harms are two sides of the same coin; researchers cannot pursue the benefits of AI without also incurring costs. But as mentioned in the overarching questions presented in the introduction to this syllabus, the question of to whom those benefits accrue and who will bear the costs is deeply important when discussing the environmental implications of AI. The selected readings provide an overview of how the benefits and costs are perceived, as well as specific examples of AI uses and their impacts.

Key questions:

1. How might we conceptualize the environmental implications of AI as an ethical issue?
2. What type of framework should stakeholders develop to weigh environmental considerations during the design, development, and implementation of AI?

- Vinuesa, R., Azizpour, H., Leite, I. et al. (2020). "The role of artificial intelligence in achieving the Sustainable Development Goals." *Nature Communications* 11: 233. https://doi.org/10.1038/s41467-019-14108-y
    - This piece by Vinuesa et al. considers the potential environmental implications of AI in the specific context of the UN's 2030 Agenda for Sustainable Development, showing evidence that AI may act as an enabler of achieving a majority of the goals. However, the authors also find that AI development may inhibit the delivery of 35% of the target goal. As noted in many other readings selected for this syllabus, this article highlights the need for regulatory insight and oversight of AI-based technologies, and that a failure to provide oversight could lead to gaps in transparency, safety, and ethical standards.

- Brevini, B. (2020). "Black boxes, not green: Mythologizing artificial intelligence and omitting the environment." *Big Data & Society.* https://doi.org/10.1177/2053951720935141
    - Like other emerging technologies, AI has been touted as the solution to many of society's problems. Brezini is interested in engaging with the costs that are often ignored in those idealizations – particularly the environmental costs. She first reviews the promised benefits before detailing how AI's energy use, waste production, and pollution, as well as how AI has become critical to apparatuses and technologies that deplete scarce resources. With this foundation, Brezini argues that environmental costs should be not only be included in all AI development processes, but that they should be centered in that work.

- Crawford, K. and Joler, V. (2018). Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources. *AI Now Institute and Share Lab*. Retrieved from: https://anatomyof.ai.
    - Crawford and Joler create an "anatomical map" of the scale of the system that works in support of human-machine interactions, like when using a product such as the Amazon Echo. The point of this mapping is to capture the vast matrix of capacities invoked – i.e., "interlaced chains of resource extraction, human labor and algorithmic processing across networks of mining, logistics, distribution, prediction and optimization" – because the scale of the system is beyond human cognitive ability. Crawford and Joler's goal is only partly to demonstrate how small conveniences performed by Echo actually represent a use of resources of an order of magnitude greater than the human energy and labor that could have been exercised in that moment (e.g., turning on the light oneself). More importantly, their mapping is an important first step in helping us to if we to grasp the underlying systems so that we can better govern technical infrastructures.

- Strubell, E., Ananya Ganesh, and Andrew McCallum. (2019). "Energy and Policy Considerations for Deep Learning in NLP." In the 57th Annual Meeting of the Association for Computational Linguistics (ACL). https://arxiv.org/abs/1906.02243
    - Using the training of neural networks for natural language processing (NLP) as an example, Strubell et al. emphasize how the models are costly to train and develop, both financially and environmentally. Any gain in NLP accuracy is dependent on the availability of exceptionally large computational resources, which in turn require substantial energy consumption. The authors' objective is to bring the environmental costs to the attention of NLP researchers. In addition, they propose recommendations to "reduce costs and improve equity in NLP research and practice."

- Spelda, P. and Stritecky, V. "The future of human-artificial intelligence nexus and its environmental costs." *Futures* 117(102531): 1-5. https://doi.org/10.1016/j.futures.2020.102531
    - Spelda and Strictecky provide a technical example of how experimenting with machine learning and AI can lead to significant environmental costs (due to carbon emissions) for little technological progress. Their "techno-philosophical way of thinking" is presented as a counterpoint to more "normative" approaches for how to frame the environmental implications of AI as an ethical issue.

Future of Work

The growing list of AI applications that surpass human performance raises a fundamental question of which tasks will be performed by human in the future. Evaluating human's uncertain future roles creates ethical questions of agency, responsibility, and the role of labor within a national economy.

Key questions

1. What underlying ethical principles capture the tradeoffs of increased automation of labor?
2. Given the uncertainty around the future of work, and a history of incorrect predictions, how should policymakers approach the automation of labor?
3. Paint a dystopian and utopian future of a highly-automated world. What ethical concerns are created and what contemporary ethical issues are potentially resolved in either scenario?

Sources:

- Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation?. *Technological forecasting and social change*, 114, 254-280. https://www.sciencedirect.com/science/article/pii/S0040162516302244?via%3Dihub
    - Though more technically than ethically framed, this widely-discussed paper can help students examine the methods of how the impact of automation is assessed and the uneven distribution of the expected automation of work. It may also be worth looking at critical responses to this work for a discussion, though many responses focus on technical disputes.

- Keynes, J.M. (1933). Economic possibilities for our grandchildren (1930). *Essays in persuasion*, pp. 358–73.
    - Grounding a conversation in Keynes brief but canonical essay on automation helps to consider the core ethical issues while also recognizing the need for intellectual humility in contemporary predictions of the future.

- Manyika, J., & Sneader, K. (2018). *AI, Automation, and the Future of Work: Ten Things to Solve for*. McKinsey & Company. https://www.mckinsey.com/featured-insights/future-of-work/ai-automation-and-the-future-of-work-ten-things-to-solve-for#
    - This report provides an overview of the extent of automation in the workforce as well as listing 'Ten Things to Solve For', which can serve as a useful starting place for discussing ethical concerns underlying the ethics of automating work.

- McKay, C., Pollack, E., & Fitzpayne, A. (2019). *Automation and the Changing Economy, Part II: Policies for Shared Prosperity*. The Aspen Institute: Future of Work Initiative. https://assets.aspeninstitute.org/content/uploads/2019/04/Automation-and-a-Changing-Economy_Policies-for-Shared-Prosperity_April-2019.pdf

  - This second half of an Aspen report lays out policy interventions in anticipation of the future of work. What are the ethical concerns underlying the policy interventions? What problems are they trying to solve, and on whose behalf?

- Osoba, O. A., & Welser, W. (2017). *The risks of artificial intelligence to security and the future of work*. RAND Corporation. https://www.rand.org/pubs/perspectives/PE237.html
  - Osoba's and Welser's short piece provides a helpful framing of the concerns about automation, introduces a range of topics, and integrates the concerns of experts from a wide range of methodologies

Rights for Robots

The authors of each of the selected readings note that their work is a contribution to a larger and ongoing conversation on the rights of robots and the ethical considerations of developing robots destined for human interaction. The readings highlight different threads of that discussion, ranging from whether robots can be considered "human" in any context to whether designing robots with moral competence can mitigate some of their potential harms. Though the readings provide critical engagement and responses to these questions, the literature shows that consensus on how to respond is still not yet settled.

Key questions

1. Are robots human? Could they become human?
2. Should we have conversations now about the potential, abstract possibilities of future AI/technological developments? (i.e., Current robot capabilities are nowhere near a future vision in which they are "some type of social partner".)
3. What are the ethical questions we must answer in order to develop morally competent use of robots in society? And in which order should we address them?
4. What are the rights of humans who interact with robots (e.g., children who are watched by robot nannies)?

Readings

- Asaro, Peter M. (2006). "What should we want from a robot ethic." *International Review of Information Ethics,* 6 (12): 9-16. https://philpapers.org/rec/ASAWSW
  - In this article, Asaro takes the position that it is worthwhile to develop a coherent framework of robot ethics that can cover all possible outcomes of robot development, including the possibility that they might one day have fully autonomous moral agency. Asaro argues that if we think about these issues from the perspective of legal responsibility, we are more likely to arrive at practical answers. He believes that legal requirements will compel robotics engineers to build ethical robots, and it will provide a practical system for understanding agency and responsibility. And, legal theory should provide a means of thinking about the distribution of responsibility in complex sociotechnical systems."

- Gunkel, David J. (2018). "The Other Question: Can and Should Robots Have Rights?" *Ethics and Information Technology*, 20: 87-99. https://doi.org/10.1007/s10676-017-9442-4
    - Gunkel takes a philosophical approach to understanding the question of whether robots *can* and *should* have rights? He first engages with the "is/ought" problem of the question, before identifying and assessing four modalities concerning social robots and rights in the existing literature on robot rights. Gunkel then proposes an alternative configuration of how to conceptualize this question, raising the issue of what else may be included in an expanded thinking about rights.

- Birhane, A. and van Dijk, J. (2020). Robot Rights? Let's Talk About Human Welfare Instead. *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. February: 207-213. https://doi.org/10.1145/3375627.3375855
    - Birhane and van Dijk closely review the "robot rights" debate, the question of "robot responsibility", and the polarized responses to these issues in this article. Although Birhane and van Dijk argue that robots should be denied any rights, they do not seek to contribute to the debate in its current framing. Rather, they argue that at its base, the debate over "robot rights" is improper, as robots inherently cannot be considered as equivalent to humans. Instead, robots are to "mediators of human being". They call for a refocusing of the debate to engage with other, urgent ethical concerns, such as machine bias, machine-elicited human labor exploitation, and the erosion, and how these issues impact society's least privileged individuals.

- Malle, Bertram F. (2016). "Integrating robot ethics and machine morality: the study and design of moral competence in robots." *Ethics and Information Technology*, 18(4): 243-256. https://doi.org/10.1007/s10676-015-9367-8
    - In addition to engagement with the questions of how humans should design, deploy, and treat robots, Malle examines the type of moral capacities a robot should have and how those capacities should be implemented. Malle's perspective is that designers/developers need to commit to building morally competent robots, and that, if they do, it will be an important step towards resolving some of the ethical concerns over robots in society.

- Sharkey, Noel and Amanda Sharkey. (2010). "The Crying Shame of Robot Nannies: An Ethical Appraisal." *Interaction Studies* 11(2): 161-190. https://philpapers.org/rec/SHATCS
    - Sharkey and Sharkey provide an overview of research on and development of childcare robots, as well as a close look at how childcare robots work in real-life settings and the concerns that arise from those roles. In the context of the other selected readings, this article provides a detailed, tangible example of human-robot interactions. And it raises important questions about the rights of

humans who must interact with robots even if they have not chosen or consented to those interactions.

Annex 3

Pardee RAND Graduate School Ethics Hackathon:

Ethics of a COVID-19 Data Platform

1. Concept and Background

While the concept of hackathons has traditionally referred to cybersecurity competitions, 'policy hackathons' have also emerged. These hackathons are typically team-based competitions that involve activities such as proposing novel policy solutions to a vexing social dilemma[24] or analyzing a dataset with policy relevance for insights and to create helpful visualizations.[25]

This memo documents the Pardee RAND Graduate School's (PRGS) first *Ethics Hackathon*, which was made possible through support from the PRGS Tech and Narrative Lab (TNL) and the Public Interest Technology University Network (PIT-UN). PRGS previously executed two technically-oriented hackathons focused on Dark Web activity and opioid treatment data.[26] The launch of the ethics hackathon provided an opportunity to build off the success of previous hackathon efforts while more fully integrating ethics throughout the PRGS's curriculum, a key aspect of the school's recent redesign.[27]

What distinguishes an ethics hackathon from traditional policy hackathons is the centrality of ethical concerns. Rather than encouraging students to design policy 'solutions' for specific outcomes, the focus is on practically considering the ethical implications, such as potential harms to vulnerable groups, of a specific program or technical tool. A core tenet of PRGS's TNL is to consider both the applications and implications of new technology, and the ethics hackathon offers a structured format for students to consider these issues.

2. Design

This hackathon focused on a new platform designed by the COVID Alliance,[28] a nonprofit organization building a research collaboration platform for COVID-19 response that includes access to multiple national datasets. The COVID Alliance platform seeks to provide live geolocation data, hospital capacity data, infection data, and other information to U.S.

[24] Results include MIT's Institute for Data, Science and Society's 2018 Policy Hackathon focused on making Boston carbon-neutral by 2050 https://issuu.com/policyhackathon/docs/mit_policy_hackathon_proceedings_fi; and Stanford's Institute for Economic Policy Research's 2019 hackathon focused on improving prisoner reentry in California https://siepr.stanford.edu/events/sig-siepr-policy-hackathon

[25] A recent example include Penn State's COVID-19 hackathon from publicly available datasets https://penntoday.upenn.edu/news/covid-19-hackathon

[26] Pardee RAND Graduate School's Tech and Narrative Lab's 2019 hackathon analyzing sales from the dark https://www.prgs.edu/news/2019/dark-web-hackathon.html; and 2018 hackathon analyzing opioid treatment data https://www.prgs.edu/news/2018/opioid-hackathon.html

[27] https://www.prgs.edu/degree-program/redesign.html

[28] https://www.covidalliance.org/

decision-makers to enable public health interventions. Several PRGS students are members of the COVID Alliance and initiated the collaboration.

The hackathon offered a mutually beneficial partnership for the COVID Alliance and PRGS students. The COVID Alliance would receive feedback on the platform related to ethical considerations they may have not considered. In turn, PRGS students would gain practical experience advising a receptive partner while demonstrating their broad analytical skillset.

The Ethics Hackathon included a dozen student participants, along with 4 faculty advisors, divided into 3 teams. The work occurred over 4 weeks and involved an initial kickoff meeting, weekly office hours with the COVID Alliance, and a final meeting where teams showed the results of their efforts.

The primary question posed to each team was: What are the potential ethical harms of the COVID platform and how can they be mitigated through design features, visualizations, and use-restrictions? This question was divided into 2 specific tasks:

1. Demonstrate potential harms of data bias.
    a. Broadly reflect on the short- and long-term ethical implications of unrepresentative data (e.g. location/mobility data, test results data, etc.) and how the platform might contribute to inequity.
    b. Build and describe an illustrative scenario demonstrating at least one specific ethical harm.
2. Propose and demonstrate actionable recommendations to address **at least one** of the following questions:
    a. What statistical methods should be deployed to identify and improve data representativeness issues on this platform? How would these statistical interventions address the inequity identified in Task 1?
    b. What additional data, features, information presentation principles, terms of use, or other ideas should be integrated into this tool to address the inequity identified in Task 1?

3. Outcomes

At the end of the Ethics Hackathon each student group presented their findings to their peers and faculty advisors. While the groups' emphases varied, the specific ethical concerns rested in three main categories:

(1) <u>Data representation concerns</u>: how representative are the underlying data signals on the platform;
(2) <u>Data privacy concerns</u>: how re-identifiable are users from their data on the platform; &
(3) <u>Data use concerns</u>: how would the data be interpreted and subsequently used by policy-makers or others.

Concerns about the underlying data focused on analyzing the platform's use of geolocation data based on cell phone pings that potentially fails to accurately represent various populations. Overall, concerns focused on how the geolocation data sources rely on unequally distributed hardware and network characteristics. For example, teams observed that lower income individuals and elderly individuals both demonstrate lower levels of smart phone usage and are thus underrepresented in the dataset. Teams also observed that rural populations' phones have fewer pings and data transactions, in part due to less reliable cell phone service. These elements of data underrepresentation were not considered insurmountable, and groups presented various methods to analyze and try to compensate for gaps in the dataset, including comparisons against census tract data as a trustworthy baseline. Other suggested methods to compensate cell phone data for areas with lower device penetration involved weighting regions to create a balanced 'pings per capita.'

Overall, the groups' findings about the underlying data quality highlighted the importance of couching technical questions of how the data might be used with skepticism over who the data represents. While the groups' findings did not mean that the data could not be used to inform valuable insights, they pointed out the need to ensure that the limitations and potential inequity associated with the data are addressed. This issue was considered especially important given the public health interventions potentially being informed by data pulled from the platform.

In addition to concerns about the representativeness of the data itself, groups also highlighted how ethical concerns hinge on knowing who the users of the data are and what types of interventions they intend to implement. Teams noted that questions of whether the data was sufficiently representative depended on what subset of the data was being considered, and for what ultimate purpose. For example, local decisionmakers focused on smaller regions might need to be more wary of representation issues exacerbated by smaller sample sizes. There were also concerns over interpretation of the data by end-users, particularly around the correlation between data availability and factors such as race and wealth, especially if the tool is used for more coercive policy such as ordering quarantines.

One specific example referenced a recent paper proposing targeted neighborhood lockdown in NYC based on movement from geolocation data.[29] The group demonstrated how the analysis failed to consider the equity implications of that policy, namely ways in which targeted lockdowns might disparately harm lower-income populations who might be less able to complete their work remotely and were engaged in necessary movement.

Another concern with the platform was the problem of data privacy and the possible reidentification of individuals. These concerns are familiar in contexts where geolocation datasets are used, but groups highlighted the risk that the use of cell phone location data potentially put the identify of individuals at risk. Given this risk, it is important to verify appropriate use of such sensitive data.

Suggested Ways Forward based on Hackathon Findings:

To address these issues, the COVID Alliance will need to clearly communicate the limits of the data and establish rules clarifying acceptable and unacceptable use. While specific recommendations on effective safeguards require more specific knowledge of the end users of the tool, the combined efforts of the hackathon groups suggest the creation of an 'ethical checklist' of features and considerations that could be used to inform the analysis and interpretation of data in the platform. Issues or potential concerns could be articulated for each phase of analysis and decision-making and ensure well-intentioned users of the platform consider the limitations of their thinking. Potential items could include:

[29] https://bfi.uchicago.edu/wp-content/uploads/BFI_WP_202057-1.pdf

Data Representation

- Have you verified the underlying dataset is accurately representative along various demographic characteristics such as:
  - Age
  - Ethnicity
  - Income
  - Race
  - Population density
- Have you compared your population sample to external data sources to ensure its accuracy?
  - I.E. Does the population/area (as determined by Census population/FIPS code) match the device-or-User_ID/area (from partner databases)?
- Have you ensured the data cannot be used to reidentify individual's identity?

Data Interpretation and Use

- Does your analysis explicitly state what types of interventions this research is and is not intended for?
- Have you articulated sensitivity analysis to explicitly state the limits of certain parameters?
- Have you considered the risks to lower-income or minority communities of coercive policies that involve quarantines or lock downs?
- Can the results of your analysis be politicized? Do you find differences in quarantine compliance across racial or economic lines?

While no ethical checklist would single-handedly eliminate misinterpretation of data or mitigate all risks of misuse or abuse, an explicit listing of considerations could be useful in minimizing harm. Such a checklist could also be useful in socializing techniques and ethical best practice across researchers, which is particularly valuable in cross-disciplinary efforts involving researchers trained by different standards and focuses.

4. Lessons for Future Ethics Hackathons

PRGS's experience hosting an ethics hackathon proved the validity of the concept, though was not without improvement for future iterations. The factors contributing to and limiting the hackathon's success provides a set of lessons for both PRGS and similar institutions as they host future ethics hackathons.

The success of the hackathon resulted from a cooperative working relationship with the external client, student enthusiasm to participate, and the diversity of hackathon findings. The COVID Alliance was open and amenable to student feedback and were happy to share their time with students as they worked with the PMP. This relationship was certainly aided by the representation of PRGS students and alumni within the COVID Alliance, but similarly positive relationships could certainly be achieved without overlapping personnel through clear communication and expectation management. Student enthusiasm most likely stemmed from PRGS's culture of collaborative experimentation, positive engagement with previous hackathons, and partial compensation for students' time through the Tech and Narrative Lab. Lastly, the diversity of findings was a result of crafting multidisciplinary student teams. While the PRGS student body pulls from a range of backgrounds and naturally results in diverse student teams, similarly diverse groups could result from collaboration between multiple academic departments.

Limitations to student success stemmed from technical difficulties with getting onboarded to the platform, and potentially from the broad scope of the required deliverable. There is an inherent tradeoff to evaluating a novel technical platform; while feedback in the early stages of a tool can more heavily influence its development, it also increases the likelihood of unprecedented technical issues. Onboarding the PRGS students to the platform hit several delays, all of which were attended to by the COVID Alliance volunteers, but those delays created a gap between the kickoff meeting and student engagement with the tool.

Future ethics hackathons should seek to ensure an immediate and continuous working environment for participants as much as possible. Due to unrelated work interruptions, student groups did not have as much time as intended to develop their ethical considerations into usable frameworks, and more time could have yielded more robust results.

PRGS is eager to develop and run additional ethics hackathons. Ethical assessments of technical tools will certainly continue to be necessary, and ethics hackathons offer an opportunity to evaluate emerging technology while developing the skills of students.