Evidence-based AI Ethics

by

William Boag

B.S., University of Massachusetts, Lowell (2016)
S.M., Massachusetts Institute of Technology (2018)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

© Massachusetts Institute of Technology 2022. All rights reserved.

Author . . . .
Department of Electrical Engineering and Computer Science
May 13, 2022

Certified by . . . .
Peter Szolovits
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . .
Leslie A Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Evidence-based AI Ethics

by

William Boag

Submitted to the Department of Electrical Engineering and Computer Science
on May 13, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

With the rise in prominence of algorithmic-decision making, and numerous high-profile failures, many people have called for the integration of ethics into the development and use of these technologies. In the past five years, the field of "AI Ethics" has risen to prominence to explore questions such as "how can ML algorithms be more fair" and "what are the tradeoffs when incorporating values such as fairness or privacy into models." One common trend, particularly by corporations and governments, has been a top-down, principles-based approach for setting the agenda. However, such efforts are usually too abstract to engage with; everyone agrees models should be fair, but there is often disagreement on what "fair" means. In this work, I propose a bottom-up alternative: Evidence-based AI Ethics. Learning from other influential movements, such as Evidence-based Medicine, we can consider specific projects and examine them for "evidence." We draw from complementary critical lenses, one based on utilitarian ethics and one from intersectional feminism to analyze five case studies I have worked on, ranging from automatically-generated radiology reports to tech worker organizing.

Thesis Supervisor: Peter Szolovits
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

In 2021, Jo Melville graduated from MIT's Chemistry program. In his dissertation, he aptly observed: "It occurs to me that there exists in theory some maximum length that is socially acceptable for a thesis acknowledgments section — a length which is, perhaps, drastically exceeded here. For this I make no apology, as the number of persons to whom I owe this degree is, far from a mark of shame, a point of distinct pride." I applaud and cheer Jo's sentiment, and in this section, I will follow suit on my many acknowledgments. This degree was not something I accomplished by alone, nor were my individual efforts the things that made my time here meaningful.

First, I must thank my amazing partner, Cindy Schilling. Thank you for keeping me sane, especially as I took on way too much this year while trying to write my thesis. I love you and I can't wait for our next of many chapters together.

Thank you to my committee: Pete Szolovits (advisor), Danny Weitzner, Catherine D'Ignazio, and Marzyeh Ghassemi. Your feedback and mentorship over the years has been incredibly valuable both to this thesis and to me as a person.

I have met so many impressive and inspiring friends during my time at MIT. Grad school is hard enough on its own, and then add in a global pandemic, and it's impossible to overstate how important it was for us all to take care of each other. A few specific shoutouts:

- Ajay Brahmakshatriya for being my CSC partner in crime for the last 3 years (2019 CSAIL Olympics, numerous CSAIL karaokes, trading CSC exec board positions, handling the CPGSC).
- Becca Black for your unrivaled empathy and great opinions on everything (e.g., science policy, the Fast and Furious franchise, the Ezra Klein Show)
- Irene Chen for being one of the most impressive people I've ever met. The only time I ever had imposter syndrome was when I compared myself to you in Fall 2019, because you seem like you can do it all, and you make it look easy. You were an amazing co-TA, and I bring out the best in myself when I try to get on your level.

- Matthew McDermott for your thoughtfulness, rigor, and curiosity. When I started at MIT, I treated being a meat eater as part of my identity. But after I came to greatly respect you, and I learned that you were a vegan, it pushed me to reflect on my assumptions and defense mechanisms, and I've been a vegetarian for 4 years now and counting.
- Haley Schilling for both introducing me to Cindy and for suggesting the "evidence-based" framing for uniting the different case studies in this thesis.

Additionally, thank you to Divya Shanmugam, Serena Booth, Amanda Beck, Bianca Lepe, Maz Abulnaga, Will Stephenson, James Minor, Akshay Narayan, Jonathan Behrens, Joana Matos Fonseca da Trindade, Hassan Kané, Micah Smith, Flora Meng, Leilani Gilpin, Christina Ji, Jenn Wang, Harini Suresh, Michael Specter, Jonathan Frankle, Julius Adebayo, Cecilia Testart Pacheco, Heather Berlin, Maggie Makar, Tarfah Alrashed, Sam DeLaughter, Sarah Cen, Martin Nisser, Mark Hamilton, Laura Brandt, Derek Lueng, Geeticka Chauhan, Rohan Chitnis, Emily Alsentzer, Davis Blalock, Liane Bernstein, Ella Wassweiler, Peter Satterthwaite, Kruthika Kikkeri, Sarah Muschinske, Liane Makatura, Zhi Xuan Tan, Anna Zeng, Katie Matton, Maddie Laitz, Gab Nahimias, Lucy Hu, Katherine Taylor, Diana Gong, Adam Trebach, Lucky Pattanaik, Rebecca Spiewak, Claire McKenna, Tej Chajed, and Monica Agrawal.

Thank you to my roommates over the years: Taylor Matte, Joe Bruni, Michael Noseworthy, and Justin Bell.

Thank you to my friends outside of MIT: Dom Geehan, Alyssa McCallion, my Patriots friends, Maggie Regan, Andrew Verras, Neil Morrissey, supporters of Dracut DI, and more. Special shout out to Kate Marston for helping me grow as a person, and for being an inspiration to me by overcoming so much to go back to school.

Thank you to all of my fellow student-worker advocates fighting to make MIT its best self. Being part of the group of students working on the 2020 EECS DEI petition (with Ajay, Anna, Liane, Husayn, Peter, and Sarah) is one of my proudest accomplishments. Special shout outs to Caris, Xuan, Maddie, and Peter for being some of the most dedicated and inspiring people I've worked with. Thank you to all of my union friends, from our early days in a small room to the RISE campaign to winning the election. Especially to Madeleine Daepp.

Two student groups have been very important to me during grad school: CSC and SPI. Thank you so much to Sara Achour and Jon Gjengset for being such welcoming leaders (Sara, especially, as President) when I started at MIT. I can't fully express how much the two of them shaped how much CSC means to me and how important it has been to me that I carry on that legacy and be welcoming to others. For SPI, I've loved finding and helping grow the science policy community at MIT! Thank you to Micah Smith for introducing me to it, to Serena Booth for being an incredible VP to me the year I was President, and to Kate Stoll for your constant support both of SPI and of my individual professional development.

Thank you to the staff, faculty, and others who have been supportive of me and for so many others, including Leslie Kolodziejski, Janet Fischer, Fredo Durand, Katrina Lacurts, Daniel Jackson, and more. These people have helped generations of students before me and they will help generations of students after me. I am proud to have met and worked with them.

Thank you to some role models that I've never actually met, including Tom Brady, Ezra Klein, Elena Kagan, Charles Houston, Stephen Breyer, Nancy Pelosi, and Barack Obama.

Thank you to the IPRI community, especially Frankle and Grace for helping me feel welcome and part of the community in early 2019 when I was trying to get more involved. Taking and then TA'ing 6.805 was one of the best experiences I had at MIT. Especially TA'ing for Hal Abelson, who was a role model of mine before I'd ever dreamed of getting into MIT. Getting to work with him has been such an honor.

Thank you to my MEDG community, which has been a really special place to be for the last 6 years: Matthew, Elena Sergeeva, Di Jin, Wei-Hung Weng, Geeticka, Emily, Heather, Harry Hsu, Sam Finlayson, Eric Lehman, Fern Keniston, Marzyeh Ghassemi, Tristan Naumann, Franck, Jenny, Michele, Joao, Ziyu, Hassan, Harini, Tiffany, and Phoung. I became the senior PhD student partway through my 2nd year, and I didnt know what I was doing at first, but I think with practice I was able to rise to the occasion.

Thank you to ALL of my co-authors: Hassan Kané, Saumya Rawat, Alex Goehler, Catherine D'Ignazio, Harini Suresh, Bianca Lepe, Olga Kovaleva, Tom McCoy, Leo Celi, Anna Rumshisky, Pete Szolovits, Roy Perlis, Wei-Hung Weng, Marzyeh Ghassemi, Shems Saleh, Harry Hsu, Emily Alsentzer, Dustin Doss, Tristan Naumann, Lauren Erdman, Guanxiong Liu, Matthew McDermott, Gabi Berner, Jack Murphy, Bret Nestor, Kaisa Elomaa, Di Jin, Michael Hughes, Anna Goldenberg, David Sontag, Pravin Kamble, Irene Chen, Song Wang, Michelle Luo, Heather Berlin, Avital Baral, Mayukha Vadari, Saurabh Kulshreshtha, Elena Sergeeva, Joey Liu, Yuan Ling, Kevin Buchan, Dimeji Farri, Sadid Hasan, Michele Filannino, Ozlem Uzuner, Kahyun Lee, Kathy Lee, Vivek Datla, Ashequl Qadir, and Julie Vaughn.

Pete taught me that "credit is infinitely divisible" and when I saw Justice Breyer give a speech, he similarly said to be generous with credit because oftentimes you can choose between getting the credit and actually getting something done (and if you're successful, there will be credit to go around, whereas if you're unsuccessful, well... who wants credit for that?).

Thank you to all of the managers, mentors, and champions from the 6 internships that I had: Quentin Palfrey and Lena Ghamrawi at IDAC, Bhavya Lal Tom Colvin at STPI, Kade Crockford at the ACLU of Massachusetts, Kate Stoll from the MIT Washington Office, and Josh Israel, Zach Cross, and Farzad Mostashari at Aledade.

Thank you to all of my UROPs and MEngs that I've had the privilege of mentoring: Sim Dy (U), Gabi Berner (U), Ramya Nagarajan (U), Shelley Choi (U), Julie Vaughn (U, M), Mercy Oladipo (U), and Saumya Rawat (U, M). These experiences have taught me a lot about myself and how to try to bring out the best in others around me. Of course, it helps to have such brilliant and hard-working partners to start with. I'm proud that my UROPs have collectively gotten 4 (hopefully 5) papers published and won 2 summer research fellowships.

Thank you to my mentors who I've pestered for countless hours and/or TA'd for: Danny Weitzner, Catherine D'Ignazio, David Sontag, Hal Abelson, and R. David Edelman. You might not have seen or remember what I was like when I started here

6 years ago, but the transformation has been incredibly large. And a big part of that has been because of the people I look up to and want to be like.

Thank you to my parents for their support. Whether its picking me up because I missed the train from Boston, driving me to my summer internship in DC, or letting me come home to stay safe from the new worldwide pandemic.

And finally, thank you so much to my advisor, Pete Szolovits. You have been such an incredible mentor and role model. Your kindness, patience, wisdom, and humor have been exactly what I needed, and you gave me the space to explore my interests and research ideas, even when they seemed a little goofy to you (and still do). I remember four years ago, when my mom got into a car accident, my brother couldn't take much time off from work and my dad had a real job and could only take a couple weeks before he needed to return to directing the people who worked for him. But Pete essentially let me take 2 months off — it was a very reduced workload — as I stayed in the hospital and then moved back home to help Mom. I think that perfectly captures Pete's mentorship approach: as a result of giving me space, I spent a lot of time interacting with hospital nurses and that first hand experience helped me think about how to explore doctor-patient mistrust for my Masters thesis. And it led to be a better research project as a result. Pete isn't a carpenter that shapes exactly how his students turn out, he's a gardener that gives his students what they need to grow into their best selves. Whatever that might be. I hope that as I go forward, I can be as good a mentor to others as Pete has been to me.

Funding sources: From 2016 to 2022, my work was funded by numerous sources, including an NIH grant (NIMH 1R01MH106577), National Science Foundation Graduate Research Fellowship Program under Grant No. 1122374, TA appointments by MIT, the MIT-Takeda Program, and the MIT AI Policy Forum.

Contents

1 Introduction 19

1.1 Academic Studies . . . 20

1.2 Principles vs. Practice . . . 21

1.3 The Value of Case Studies . . . 24

1.3.1 Contextualized Knowledge . . . 25

1.3.2 Persuasion and Verifiability . . . 26

1.4 Evidence-based Medicine . . . 26

1.5 Evidence-based AI Ethics . . . 28

1.5.1 Effective Altruism . . . 29

1.5.2 Data Feminism Analysis . . . 30

1.5.3 Putting it Together . . . 33

1.6 Chapters . . . 33

2 Radiology Report Generation 35

2.1 Problem Definition . . . 35

2.2 Related Work . . . 37

2.2.1 Radiology Report Generation . . . 37

2.2.2 Deep Learning and Image Captioning . . . 38

2.3 Model Design . . . 40

2.3.1 CNN-RNN-RNN Model Architecture . . . 41

2.3.2 Clinical Coherence . . . 43

2.4 Methodology . . . 46

2.4.1 Data . . . 46

2.4.2 Models 47

2.4.3 Evaluation 48

2.5 Results 50

2.5.1 Quantitative Results 50

2.5.2 Qualitative Results 51

2.6 Project “Evidence” 54

2.6.1 Evaluating a Choice of Project 54

2.6.2 Project Impact Evaluation 61

3 Evaluating Quality of Generated Text 67

3.1 Flaws in Standard Evaluation Metrics 68

3.1.1 Inconsistent Ranking of Methods 69

3.1.2 Challenges of the Clinical Domain 71

3.2 Attempt 1: Pilot Study in Clinician Judgment 72

3.2.1 Methodology 73

3.2.2 Results 78

3.3 Attempt 2: Multi-Dimensional Evaluation 82

3.3.1 Methodology 82

3.3.2 Results 89

3.4 Project “Evidence” 91

3.4.1 Self-Critiques of Attempt 1 92

3.4.2 Self-Critiques of Attempt 2 95

4 Racial Disparities in End-of-Life Care 99

4.1 Disparate Treatments During End-of-Life 99

4.2 Building ML: EOL Risk Prediction 101

4.2.1 Data 102

4.2.2 Methods 103

4.2.3 Results 103

4.3 Upstream Intervention: ML for ACP 105

4.4 Project “Evidence” 106

4.4.1 Algorithmic Audit 106

5 Tech Worker Collective Action 115

5.1 Problem Definition 115

5.2 Background and Related Work 117

5.2.1 Accountable Algorithms 117

5.2.2 Social Movements and Labor Organizing 119

5.3 Tech Worker Campaigns: Three Case Studies 120

5.3.1 Muslim Registry (2017) 121

5.3.2 Project Maven (2018) 122

5.3.3 Face Surveillance 123

5.4 Tech Worker Actions: A Wider Analysis 124

5.5 Building Tech Worker Power 129

5.5.1 McAlevey: Organizing for Power 130

5.6 Discussion 136

5.6.1 Theory of Power 136

5.6.2 Expectations and Timeline 137

5.6.3 Tech-specific Considerations 138

5.7 Project "Evidence" 140

5.7.1 Reconstructing the Work of This Project 140

5.7.2 Increasing Awareness of Current Labor Organizing 143

6 Analyzing Differential Privacy 147

6.1 Introduction 147

6.2 Related Work 149

6.3 Methodology 151

6.3.1 Data 151

6.3.2 Generative Model 153

6.4 Measuring Utility and Risk 156

6.4.1 Utility: Model Performance 156

6.4.2 Privacy: Membership Distance 157

6.4.3 Privacy: Attribute Inference 158

6.5 Results 158

6.5.1 Utility-Privacy Curves 160

6.5.2 Closer Look: Privacy for Cancer Patients 161

6.6 Discussion 164

6.7 Project “Evidence” 165

6.7.1 Spectrum of Differential Privacy 165

6.7.2 (Multi-)Spectrum of Data Sensitivity 167

7 Conclusion 169

7.1 Who is Doing AI Ethics? 169

7.1.1 Corporations 170

7.1.2 AI Workers 170

7.1.3 Policymakers and Regulators 171

7.1.4 Think Tanks and Advocacy Groups 172

7.1.5 Media and Journalists 172

7.1.6 The Public 173

7.2 Building Norms and Policies 174

7.2.1 Policy Synthesis: AI Policy Forum 174

7.2.2 Building Norms: Privacy Watchdog 176

A Radiology Report Model Implementation Details 179

A.1 Our Model Implementation 179

A.2 TieNet Re-implementation 180

Bibliography 183

List of Figures

2-1 Synthetic example of a deidentified radiology report in MIMIC. . . . 36

2-2 Radiology report generation as an image captioning task. 37

2-3 The model for clinically-accurate report generation. 41

2-4 **Visualization of the generated report and image attention maps.** Different words are underlined with its corresponding attention map shown in the same color. Best viewed in color. 54

2-5 Number of papers citing each public diagnostic captioning dataset's paper per year. 57

3-1 Three different annotation tasks we considered for the radiologists to perform. 74

3-2 This "1-pager" document was sent to radiologists during outreach when setting up initial conversations about this project's goals. 76

3-3 An example instance of the chosen annotation task. 77

3-4 We iteratively filled in this 2-page document (1 of 2) between meetings with our radiologist collaborator. Walking through it together helped build a shared understanding with them about what was important to measure and why. 85

3-5 We iteratively filled in this 2-page document (2 of 2) between meetings with our radiologist collaborator. 86

3-6 Scoring Guidelines provided to radiologists. 88

3-7 Task presented to radiologists, where all captions are shown at the same time. 89

4-1 Distribution of risk scores for white, Black, and Asian patients. The dotted line indicates the median risk score for that group. 104

4-2 Kaplan Meier curve of patients that die post-discharge. 112

5-1 Depiction of the collective actions tagged as "ethics" in the CAIT archive. Each action was categorized into Sharp's framework. The axes of this figure do not signify quantitative meaning. 126

6-1 **Membership Distance** curves of each sensitive attribute. This notion of utility is defined in Section 6.4.1. This notion of privacy is the Membership Distance definition found in Section 6.4.2. The x-axis is Euclidian distance units and the y-axis is Mean Absolute Error. Each point's color corresponds to that dataset's privacy budget, ϵ. 159

6-2 **Attribute Inference** curves of for each sensitive attribute. This notion of utility is defined in Section 6.4.1. This notion of privacy is the Attribute Inference definition found in Section 6.4.3. The x-axis is recall from an attribute prediction classifier and the y-axis is Mean Absolute Error. Each point's color corresponds from the privacy budget, ϵ, used to generate that dataset. 159

6-3 Restating the Membership Distance (A) and Attribute Inference (B) curves for the cancer metric. Each curve is labeled with four letters (A-D, E-H) where each letter corresponds to a dataset visualized in Figure 6-5. 161

6-4 2-dimensional visualizations of the real dataset. This view is a projection of the OASIS (x-axis) and age (y-axis) dimensions. 162

6-5 2-dimensional visualizations of the 8 datasets identified in Figure 6-3. For comparison with Figure 6-4, the OASIS score is the x-axis and the age is the y-axis. 163

List of Tables

2.1 **Automatic Evaluation Scores.** The table is divided into natural language metrics and clinical finding accuracy scores. BLEU-n counts up n-gram for evaluation, and accuracy is the averaged macro accuracy across all clinical findings. *Major class* always predicts negative findings. 49

2.2 **Clinical Finding Scores**. The precision scores for each of the labels are listed and aggregated into the overall precision scores. Macro denotes averaging the numbers in the table directly and micro accounts for class prevalence. Rows are ordered by class prevelance. 51

2.3 **Sample images along with ground truth and generated reports**. Note that upper case tokens are results of anonymization. 52

3.1 Automatic evaluation metrics of baseline methods for image captioning task. 70

3.2 Linguistic characteristics of PASCAL-50S (CIDEr's annotations) and the MIMIC-CXR radiology reports. 72

3.3 Of the 199 consensus comparisons, how often would each metric rank the two reports the same way the radiologists did? 79

3.4 Top n-grams from the explanations provided by annotators for decision-making. Phrases containing stop words were removed. 79

3.5 Readability scores of the 100 ngram-generated reports vs. the 100 random-report candidates. 81

3.6 Correlations between dimensions. Each image had 4 dimensions, each dimension's 3 annotator scores were averaged into one score per dimension. 90

3.7 Correlations between dimensions, separated by annotator. "FC" stands for factual correctness, "C" stands for comprehensiveness, "P/S" stands for presentation / style, and "O" stands for overall. 90

4.1 Rates of patients who pass away within one year of discharge rates by race. This uses the MIMIC cohort of patients with a recorded post-discharge death date. 102

4.2 Model performance by race, evaluated on the cohort from Table 4.1. We report the mean across 20 runs, plus or minus the standard deviation. 103

4.3 1-year mortality rates by race for the "unfiltered" cohort. 109

4.4 Model performance by race, trained on the "unfiltered" cohort. We report the mean across 20 runs, plus or minus the standard deviation. 111

5.1 Subset of nonviolent methods enumerated by Sharp. 125

5.2 FAccT papers (2018-2021) whose abstracts use the employment and collective terms developed by the creators of the Collective Action in Tech archive. 144

Chapter 1

Introduction

Ethics is a branch of philosophy that tries to understand right and wrong. It seeks to formulate whether given actions should be forbidden, permissible, or imperative. In 1980, Langdon Winner wrote "Do Artifacts Have Politics" which explored whether a new technology (i.e., something which changes the way a process is done to be faster, cheaper, more efficient, less complex, etc.) inherently has a political orientation [307].

Within the last decade, high profile failures of increasingly popular machine learning (ML) systems have led to another iteration of the fundamental conversation around technological progress and its potential harms. Technologists, lawyers, ethicists, and others have greatly increased the effort to study these problems in recent years. Virtually everyone agrees that things should be "ethical," but there seem to be large disagreements on what "ethical" means. There is no single authority, instead there is an ecosystem of actors, including academics, corporations, policymakers and regulators, the public, media and journalists, think tanks, and more.

Within this ecosystem, actors operate within power structures, including the law, public consensus, market forces, individual codes of ethics, etc. When one of those structures is very clear (e.g. legal protections for intellectual property, market forces discouraging explicitly racist algorithms, etc.) then actors are constrained. However, as new technology moves interactions from well-understood terrain to "the wild west" then powerful actors have freer reign to operate in the gray area. This could come in the form of academics doing exploitative data collection practices [102] or corporations employing dark patterns to nudge users into actions they wouldn't otherwise take [196].

1.1 Academic Studies

In this thesis, I primarily focus on how academics can participate in and contribute to ethical AI. In the conclusion, I "zoom out" to discuss the broader ecosystem of actors in which these efforts sit.

Conceptually, academics are meant to be truth-seeking scientists. They formulate and validate hypotheses to understand the world without a market-driven goal. An academic's incentives essentially boil down to being able to publish papers that are influential to other academics (as measured by citations, peer feedback, tenure, grants, etc.).

There are many academic fields exploring tech+ethics issues, including work in Sociology, Geography, Media and Communication Studes, Critical Data Studies, and more. One prominent academic community for AI Ethics is the algorithmic fairness conference Fairness, Accountability, and Transparency (FAccT). In 2018, FAccT was launched to bring together "a wide array of disciplines and subfields including machine learning, statistics, measurement and security, theoretical computer science, law, policy, philosophy, sociology, and interdisciplinary work touching on many of these fields" [88]. The 208 FAccT papers generated between 2018-2021 included impactful work, including scholarship which anchored the conversation for facial surveillance bans across the US [43], improved a deployed ML model that was making racially-biased predictions for millions of patients [210], and raised concerns about the un-sustainability of large language models [22]. The 5 most-cited papers from FAccT, to date, are:

- Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification [43]
- Model Cards for Model Reporting [184]
- Explaining Explanations in AI [187]
- A Comparative Study of Fairness-Enhancing Interventions in Machine Learn-

ing [89]

- Fairness and Abstraction in Sociotechnical Systems [254]

1.2 Principles vs. Practice

With such a large ecosystem of players, there is a lot of brain power devoted to exploring the ethics of AI. Perhaps naively one might ask “why don’t we know the answers yet?” One issue is that many of the available actions available are in setting policies and cultures/norms. The situations they are applied to are often complex, where a given policy might have unintended consequences. There is no shortage of ideas that one could try, but there is a bottleneck on how an idea would play out if it tried to be executed.

Abstract principles are often deliberately written so as to be acceptable to many and paper over disagreements. When countries negotiate international standards, the language is quite vague; the Organisation for Economic Co-operation and Development (OECD) Principles for Internet Policy Making carefully negotiated recommendations like “Maximise individual empowerment” and “Promote creativity and innovation” in order to be agreeable to 34 countries [211]. Who could argue with promoting creativity and innovation? A main obstacle is that there are too many frameworks which all say very similar things, and it is unclear which ones are more useful or how the reader would come to compare them.

In the case of AI Ethics, I contend that we already have an abundance of ideas, and we need more validation of how those ideas play out in practice. As of December 2021, Algorithm Watch identifies at least 173 AI Ethics Guidelines in their global inventory. Each set of principles seems reasonable enough, but they are all slightly different and without some obvious way of identifying what approaches are working well and which ones are not. Examples of the principles include:

- **Microsoft** Responsible AI Principles
  - Fairness: AI systems should treat all people fairly
  - Reliability and Safety: AI systems should perform reliably and safely

  - Privacy and Security: AI systems should be secure and respect privacy
  - Inclusiveness: AI systems should empower everyone and engage people
  - Transparency: AI systems should be understandable
  - Accountability: People should be accountable for AI systems
- **New York Times**: Seeking Ground Rules for AI
  - Transparency: Companies should be transparent about the design, intention and use of their A.I. technology.
  - Disclosure: Companies should clearly disclose to users what data is being collected and how it is being used.
  - Privacy: Users should be able to easily opt out of data collection.
  - Diversity: A.I. technology should be developed by inherently diverse teams.
  - Bias: Companies should strive to avoid bias in A.I. by drawing on diverse data sets.
  - Trust: Organizations should have internal processes to self-regulate the misuse of A.I. Have a chief ethics officer, ethics board, etc.
  - Accountability: There should be a common set of standards by which companies are held accountable for the use and impact of their A.I. technology.
  - Collective Governance: Companies should work together to self-regulate the industry.
  - Regulation: Companies should work with regulators to develop appropriate laws to govern the use of A.I.
  - "Complementarity": Treat A.I. as tool for humans to use, not a replacement for human work.
- **Google**: Responsible AI Practices
  - Use a human-centered design approach
  - Identify multiple metrics to assess training and monitoring
  - When possible, directly examine your raw data
  - Understand the limitations of your dataset and model
  - Test, Test, Test
  - Continue to monitor and update the system after deployment

- **IBM**: Principles for Trust and Transparency
  - The purpose of AI is to augment human intelligence
  - Data and insights belong to their creator
  - New technology, including AI systems, must be transparent and explainable
- **Intel**: AI Policy Whitepaper
  - Increased automation should not translate to less privacy protection;
  - Explainability needs more accountability;
  - Ethical data processing is built on privacy;
  - Privacy protects who we are (how others see us and how we see ourselves);
  - Encryption and de-identifcation help address privacy in AI.
- **Philips**: Five guiding principles for responsible use of AI in healthcare and healthy living
  - Well-being
  - Oversight
  - Robustness
  - Fairness
  - Transparency
- **Sony**: AI Engagement
  - Supporting Creative Life Styles and Building a Better Society
  - Stakeholder Engagement
  - Provision of Trusted Products and Services
  - Privacy Protection
  - Respect for Fairness
  - Pursuit of Transparency
  - The Evolution of AI and Ongoing Education

Many of these seem like fine ideas, but what does it mean to care about fairness, reliability, safety, privacy, security, inclusiveness, transparency, and accountability all at once? What happens in a scenario where privacy and fairness might be in tension with each other, such as wanting to audit an algorithm which requires collecting additional data? Facebook claims[1] to have "Five Pillars of Responsible AI" which are: Privacy and Security, Fairness and Inclusion, Robustness and Safety, Transparency and Control, and Accountability and Governance. However, after a series of scandals, it does not seem like they're living up to their stated values of privacy [295], security [172], fairness [72], or safety [262].

Bamberger and Mulligan [18] studied corporate privacy, and observed the difference between "privacy on the books" (i.e., what the official rules are) and "privacy on the ground" (i.e., what is done in practice). This distinction similarly would apply to these other values as well, even if they have not all been invesitgated as closely as privacy practices have.

There is a need for examining not just what one says but what they do.

1.3 The Value of Case Studies

In contrast to abstract frameworks, case studies demonstrate how the competing tensions of different values get resolved. Sometimes actors assert they can achieve multiple values without a seemingly necessary trade off, but a case study analysis shows whether that is true "when the rubber meets the road." Additionally, case studies are more likely to be useful even if the reader has different values than the author, because the value-add is not from the answer but from demonstrating what is interesting about the context of the decision. When frameworks are too abstract so as to remove that context, they become less useful because there is not guidance in how (let alone why) to resolve when different values collide.

To illustrate the benefits that case study analysis can contribute to decision-making, I will use a field my reader is less likely to have much experience or knowledge in: education policy. This will help simulate how policymakers — with their wide portfolio of issues to legislate for — often must navigate areas they do not have much expertise in. I demonstrate how case studies help adjudicate disputes where either side makes a plausible-sounding argument.

[1] https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai

1.3.1 Contextualized Knowledge

Another value of case studies is that they allow one to resolve empirical questions.

For example, the Obama administration upset some of their ideological allies with their "Race to the Top" program, which incentivized states to change their education policy, including with reforms which weakened teachers' unions [41]. The policies in question did things like expand standardized testing and allow principals to fire bad teachers more easily. The administration argued that this would help identify which teachers were hurting students (especially low-income and minority students) and allow principals to rectify the situation. Teachers unions argued that standardized tests were ineffective tools for measuring educational outcomes and that these rules were a corporate-supported attack on unions and the middle class [139].

In the example above, one side says standardized testing will help students and the other says that it won't. Who is right? In 2009, the teachers unions (typically "anti-testing") worked with the Bill and Melinda Gates Foundation (typically "pro-testing") to run a study in 4 districts to pilot programs to study the impact of education policy reforms. They found that it is not as simple as "tests are always good" or "tests are always bad" but instead that these reform efforts depend on a lot of other factors, including "a trust relationship between union and district leaders; a joint focus on problem solving and learning together; teacher input in program design and implementation; voluntary participation by teachers; flexibility in program design; and an overall comprehensive approach to the entire effort" [268].

This more nuanced understanding of what solutions will or will not work allows for the parties to work towards compromise which respects each of their goals: education reformers want to ensure racial education gaps are measured and closed whereas teachers unions want to make sure they have the power to protect their members from unfair discipline and termination. This consensus-build and persuasion effect is another benefit of case studies.

1.3.2 Persuasion and Verifiability

Often times, people disagree because they are talking past each other: they can each make a plausible argument with enough anecdotes supporting their beliefs. Because pilots and case studies are able to reality-test those arguments, decision-makers end up better equipped to understand the actual impacts of a given decision. Sometimes, the results might even be surprising (e.g., if an education reform advocate thought standardized testing made intuitive sense but didn't appreciate the importance of program implementation). Those instances of unexpected results to empirical questions can be an effective way for opposing sides to actually come to an agreement because often times people are not receptive to abstract arguments based on first principles if they don't already agree with the conclusion.

In 2019, then-Presidential candidate Kamala Harris put forth a proposal which both the education reformers and the teachers' unions were happy with that proposed increasing teacher salary in order to retain the best teachers and foster a relationship of trust between schools and teachers [253]. Because there were existing successful programs to reference, both sides were willing to trust that the proposal was not out to quielty undermine their values while merely using using the right phrases to pay lip service to them.

1.4 Evidence-based Medicine

In this section, I look at how the field of medicine has developed "evidence-based medicine" as a deliberate practice for putting principles to practice. By working with existing institutions as well as developing new ones, this effort was able to change medicine's professional culture and standards. I believe that we can learn from how this movement was able to align around a framework, and then we can apply some of those lessons to the new field of AI Ethics.

The field of medicine is thousands of years old, with roots as far back as ancient Egyptian, Greek, Chinese, and more. The modern era of western medicine began around the discovery of the smallpox vaccine in the early 1800s [51]. Medicine was notorious for “snake oil”, and the desire to professionalize the field and drive out the “quacks” motivated the 1847 formation of the American Medical Association [5]. During that time, the AMA created codes of medical ethics for practitioners and aided the passage of US laws such as the Pure Food and Drug Act in 1906.

The professionalization of medicine has led to vastly improved outcomes for many, though there is always room for improvement on the path to progress. A 2010 study estimated that only 20% of clinically significant decisions made by doctors were “[based on] any prior published data and fewer than 3% of decisions were based on a study specific to the question at hand” [65].

A recent approach towards medical progress is the “Evidence-based Medicine” (EBM) movement. EBM started around 1990 to promote best practices for clinical care. The first published instance of the term “evidence-based” was by David Eddy in a 1990 article in the Journal of the American Medical Association (JAMA) which laid out the principles of evidence-based guidelines and population-level policies [75]. At a high enough level, everyone would agree that making decisions is better when based on evidence rather than based on nothing. What distinguishes EBM from being a tautology is that it emphasizes increased reliance on up-to-date published research, particularly clinical trials. It argues that because clinical judgment, mechanistic reasoning, and authoritative opinion are less reliable, their value should be down-weighted when designing guidelines of best practices.

Because EBM has been advocating for improvements for the last 30 years, we can look back on lessons to learn. Sheridan and Julian [260] survey both successes and limitations of EBM. Some of its successes include contributing to the development of clinical guidelines, calling for full disclosure of clinical trials, and contributing to awareness of “overdiagnosis” and campaigns against “Too Much Medicine” [286]. On the other hand, it has had some limitations, such as over-reliance on clinical trials and systematic reviews to the detriment of clinical judgment and mechanistic reasoning in guidelines, encouraging a bias towards “easily measurable” risks to the detriment of other values such as patient experience and dignity, and exclusively focusing on drugs and devices to the detriment of other levers such as policy and logistics/delivery.

Randomized controlled trials are considered to be the strongest form of evidence despite shortcomings like the unrepresentative-ness of clinical trial subjects (both by demographics and by the removal of patients with comorbid diseases from trials) and a conflation between average treatment effect and individual treatment effect.

Nonetheless, there is one thing that EBM has unequivocally done successfully: align policymakers, educators, and practitioners around a framework for improvement (namely, itself). A 2009 study of medical schools in the UK found that over half of them incorporated some training of EBM practices [181]. Additionally, a primary EBM institution (the Cochrane Collaboration) harnesses the efforts of nearly 40,000 volunteers to produce over 400 systematic reviews of clinical trials per year, which had almost 4 million downloads in 2010 [260].

1.5 Evidence-based AI Ethics

I am interested in applying some of the lessons of Evidence-based Medicine successes as well as their critiques. In particular — because we cannot run randomized controlled trials for many AI Ethics questions — I am interested in further examining the question of what counts as "evidence."

To address this question, I will employ two complementary frameworks: Data Feminism and some methodologies from Effective Altruism. Effective Altruism is an attempt to embody utilitarian ethics, where practitioners try to measure and optimize which actions are doing the most good for the most number of people. Data Feminism, on the other hand, comes from intersectional feminism, and applies critical examination of power. It elevates other forms of knowing, employing ethnographies and user-centered design practices. These two approaches complement each other well because utilitarianism tries to maximize an objective, often by playing to the average, whereas Data Feminism purposefully tries to understand the marginalized members of a space.

1.5.1 Effective Altruism

Because EA cultural norms highly value transparency, most organizations publish their methodologies and decision-making processes. The principal unit of analysis is trying to understand the counterfactual of a given decision (i.e. attempting to explicitly characterize what the impact would be for each choice in a decision). Because these decisions are analyzed within their given context (e.g. "Should I donate $100 to the Humane League or the Against Malaria Foundation?"), they measure effectiveness based on the *marginal* impact of the choice (i.e. instead of measuring which organization does more good overall, they would try to understand how each organization would use that extra $100 and then determine which use results in more good).

GiveWell publishes[2] its criteria for evaluating which charities to recommend: evidence for effectiveness, cost-effectiveness, room for more funding, and transparency. As mentioned above, GiveWell evaluates charities based on the marginal impact of donating to them (i.e. room for more funding). Both their "evidence for effectivness" and "transparency" criteria demonstrate their approach to what constitutes "evidence" and how much they trust it. They published their 2012 criteria[3] for assessing any evidence (e.g. awards/recognition/reputation, testimony, broad trends in data, formal studies) based on relevant reporting effects, attribution, representativeness, and consonance with other observations. For formal studies, they have more formal criteria[4] for scrutinizing the evidence, looking for:

- Do they measure attribution with RCTs, instrumental variables, regression discontinuity, controlling for confounders, or visual and informal reasoning?
- What are the likely motivations and hopes of the authors?
- Is the paper written in a neutral tone? Do the authors note possible alternate interpretations of the data and possible objections to their conclusions?
- Is the study preregistered? Does it provide a link to the full details of its analysis, including raw data and code?

[2] `https://www.givewell.org/how-we-work/criteria`

[3] `https://blog.givewell.org/2012/08/17/our-principles-for-assessing-evidence/`

[4] https://blog.givewell.org/2012/08/23/how-we-evaluate-a-study

- How many outcomes does the study examine, and which outcomes does it emphasize in its summary?
- Is the study expensive? Were its data collected to answer a particular question?
- What is the effect size and p-value of the intervention?

In general, they believe "the property of being an RCT is probably the single most encouraging (easily observed) property a study can have" although "it's possible that if preregistration were more common, [they'd] consider preregistration to be a more important and encouraging property of a study than randomization."

Relatedly, William MacAskill (an originator of Effective Altruism) focuses on identifying which causes to prioritize studying [174], as opposed to which charities are effective within a given cause. The criteria he considers are:

1. **Importance**: How useful would it be to the world if this problem was solved?
2. **Tractability**: How much of a difference can you make towards solving this problem?
3. **Neglected-ness**: How over-saturated is the space of solutions?

Finally, EA values self-criticism. For instance, GiveWell maintains a list[5] of mistakes they have made and what they learned from it, including 11 "major" issues and 25 "smaller" issues. Similarly, 80,000 Hours also maintains a list[6] of mistakes they have made and what they've learned.

1.5.2 Data Feminism Analysis

In 2020, Professors Lauren Klein and Catherine D'Ignazio published *Data Feminism*, a framework for critiquing and doing data science using concepts from the academic discipline of intersectional feminism [64]. The book challenges the perceived neutrality of data science, and treats data as a form of power. Inline with its feminist lineage, the book explores "who?" questions (e.g., data science by who? data science about who? data science using whose values?). The book is divided into 7 chapters, one per principle of Data Feminism:

[5] `https://www.givewell.org/about/our-mistakes`

[6] `https://80000hours.org/about/credibility/evaluations/mistakes`

1. **Examine Power**: “[A]nalyz[e] how power operates in the world.”
2. **Challenge Power**: “[C]halleng[e] unequal power structures and working toward justice.”
3. **Elevate Emotion and Embodiment**: “[V]alue multiple forms of knowledge, including the knowledge that comes from people as living, feeling bodies in the world.”
4. **Rethink Binaries and Hierarchies**: “[C]hallenge the gender binary, along with other systems of counting and classification that perpetuate oppression.”
5. **Embrace Pluralism**: “[T]he most complete knowledge comes from synthesizing multiple perspectives, with priority given to local, Indigenous, and experiential ways of knowing.”
6. **Consider Context**: “[D]ata are not neutral or objective. They are the products of unequal social relations, and this context is essential for conducting accurate, ethical analysis.”
7. **Make Labor Visible**: “The work of data science, like all work in the world, is the work of many hands. Data feminism makes this labor visible so that it can be recognized and valued.”

By embracing pluralism and “synthesizing multiple perspectives, with priority given to local, Indigenous, and experiential ways of knowing,” D’Ignazio and Klein [64] do not mean ‘everything is equally correct and nothing has meaning’. Feminist scholar Sandra Harding developed the concepts of feminist objectivity and standpoint theory, which challenge the notion that there is a “view from God” that can ever be neutral [113]. Instead, feminist objectivity accounts for the situated nature of knowledge and brings together multiple “partial perspectives” into a more whole understanding. In particular, standpoint theory empowers marginalized perspectives to challenge the dominant view. For instance, a vast majority of scientific studies are conducted by men using male subjects, but the results are often generalized to women [170]. In critically analyzing such studies, we would consider the standpoints of the researchers and cohort with an emphasis on the experience of the marginalized subjects. In doing so, we are able to learn from the ones who might most acutely experience the implications of a given design choice.

The third principle of Data Feminism calls for elevating emotion and embodiment. This is somewhat at odds with one of the criteria GiveWell used for evaluating studies in the Effective Altruism framework: “Is the paper written in a neutral tone? Do the authors note possible alternate interpretations of the data and possible objections to their conclusions?” Whereas GiveWell is looking for presented neutrality as an indication that the study is not putting a “thumb on the scale” towards a given result, Data Feminism would suggest that those are two separate dimensions of analysis: 1) no study is truly objective; but 2) conflicts of interest should be taken very seriously.

In adjudicating conflicts of interest, DF offers strong critiques of how money and power influence the creation of classification systems and how structures encode values. For instance, Facebook allows users to identify as non-binary on the surface but it still internally categorizes them as a binary gender for determining what advertisements to display (because their revenue comes from ads) [26]. This example illustrates that incentives inseparably shape the way systems and studies are designed.

In further considering the context of a data creation process, we are also reminded of the “paradox of exposure” in which marginalized members might prefer to remain invisible to their oppressors. For instance, in 2019, there was a Republican-led effort to add a citizenship question to the US Census [306]. Unauthorized immigrants, fearing the risk of deportation if they were counted, were less likely to complete their census, and were under-counted in the allocation of political representation and financial assistance. By explicitly considering the influence of power on marginalized data subjects, we can examine potential sources of missing data in a study. This also reinforces the seventh principle of Data Feminism: make labor visible. By uncovering some of the invisible work, you can see some of the biases of the invisible forces which generated the data.

Both EA and Data Feminism call out the difficulty in conveying uncertainty. Whereas EA scrutinizes p-values, DF looks for the ways in which the work uses tools for emotional impact, such as dynamic or stochastic visualizations. Much like how GiveWell would look for neutral tone as an indication of the work’s competence, a data feminist might keep an eye out for how effective the work is at incorporating affect into the visualization.

1.5.3 Putting it Together

Taking elements from these two frameworks, I plan to describe five case studies and analyze them to understand their impact.

Because these case studies are my own, I will be able to make the labor visible more easily than one would be able to do from an outside perspective. This can help clarify the context the work was created in and allow for a discussion of potential conflicts of interest or areas of scrutiny.

In examining what constitutes "evidence" I will consider the project's "outputs" (e.g., published papers, influence on the field, citations, number of reproductions / github forks, estimates of patients impacted), environment (e.g., estimate of likelihood similar work would be publsihed by others) and "inputs" (e.g., project team discussions, conflicts of interests, initial hypothesis).

1.6 Chapters

In this thesis, I hope to demonstrate a case study-oriented approach forward for AI Ethics. Chapters 2–6 describe five case studies in AI Ethics and analyze them with the frameworks above (Data Feminism and Effective Altruism). In the final chapter, I zoom out to discuss how the contributions of this thesis (the need for and methodology of case study analysis) fits into the broader consensus-building required for the new field of AI Ethics.

Chapter 2

Radiology Report Generation

In this chapter, I describe a case study where computer scientists provided a technical contribution to a research question. By working with clinicians to understand a problem and design a solution, this demonstrates one opportunity for the value that data scientists can bring to interdisciplinary problems. Using advanced deep learning techniques, we build a state-of-the-art radiology report generation model.

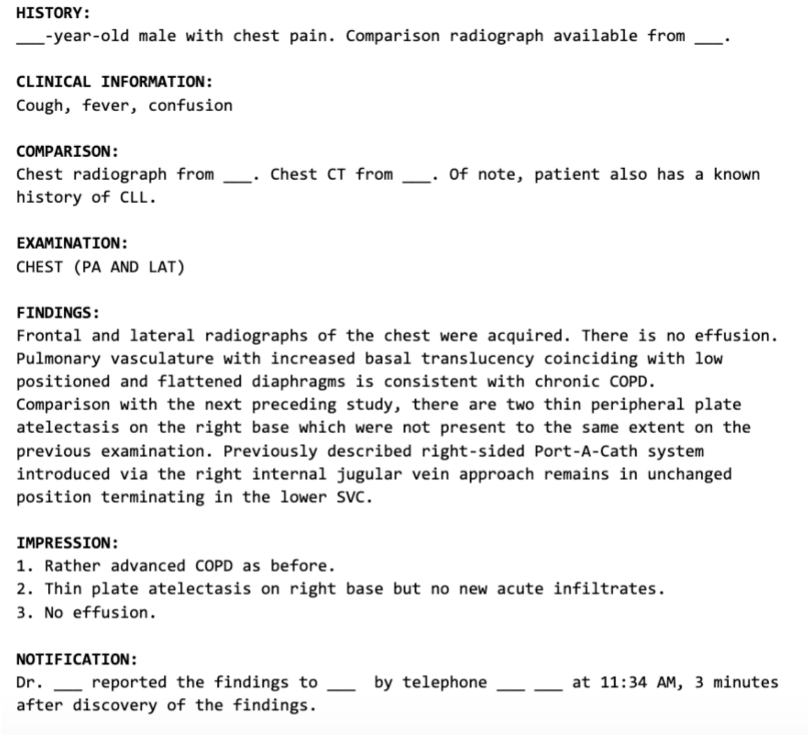
In Chapter 3, I demonstrate the importance and challenges of deeply evaluating the quality of such generated reports, which is an essential piece of the pipeline of technological innovation.

This work was done in collaboration with Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi [168].

2.1 Problem Definition

Radiology is a medical discipline that uses medical imaging to diagnose and treat diseases. Using imaging techniques like X-Rays, CT, MRI, and ultrasound, radiologists are able to measure what is ailing their patients. Typically, a patient has a non-radiologist clinician (e.g., primary care, emergency medicine, intensivist, etc) that orders radiology exams to better understand the patient's problems. A radiologist then performs the imaging and writes a report to communicate their findings to the ordering physician. These reports are often a critical piece of the patient's care,

Figure 2-1: Synthetic example of a deidentified radiology report in MIMIC.



HISTORY:
___-year-old male with chest pain. Comparison radiograph available from ___.

CLINICAL INFORMATION:
Cough, fever, confusion

COMPARISON:
Chest radiograph from ___. Chest CT from ___. Of note, patient also has a known history of CLL.

EXAMINATION:
CHEST (PA AND LAT)

FINDINGS:
Frontal and lateral radiographs of the chest were acquired. There is no effusion. Pulmonary vasculature with increased basal translucency coinciding with low positioned and flattened diaphragms is consistent with chronic COPD.
Comparison with the next preceding study, there are two thin peripheral plate atelectasis on the right base which were not present to the same extent on the previous examination. Previously described right-sided Port-A-Cath system introduced via the right internal jugular vein approach remains in unchanged position terminating in the lower SVC.

IMPRESSION:
1. Rather advanced COPD as before.
2. Thin plate atelectasis on right base but no new acute infiltrates.
3. No effusion.

NOTIFICATION:
Dr. ___ reported the findings to ___ by telephone ___ ___ at 11:34 AM, 3 minutes after discovery of the findings.

informing whether there are any diagnoses to confirm or rule out.

Like all processes involving humans, radiological practices have some variation that leads to errors and discrepancies in their findings. Limitations such as fatigue and cognitive biases can inhibit intended performance, as demonstrated by the “gorilla effect.” The gorilla effect (also called “tunnel vision” or “inattentional bias”) was named for a famous study in which 83% of radiologists were so focused on searching for lung nodules in a thorax CT that they failed to notice that an image of a gorilla (48 times the size of the average nodule) was inserted in the last case that was presented [70]. Some studies estimate radiology studies have a real-time day-to-day radiologist error rate average of 3-5% [212, 40]. These errors and discrepancies offer an opportunity for machine learning to improve upon the status quo. Although ML is not a silver bullet that will fix every problem, it does have specific qualities that it can do well, including performing tasks very quickly and performing tasks with consistency. These aspects allow ML to operate at scale, without getting tired, and with measures to mitigate known biases.

A written radiology report typically consists of sections such as *history*, *examination reason*, *findings*, and *impressions*. As shown in Figure 2-1, the *findings* section contains a sequence of positive, negative, or uncertain mentions of either disease observations or instruments including their detailed location and severity. The *impression* section, by contrast, summarizes diagnoses considering all report sections above and previous studies on the patient.

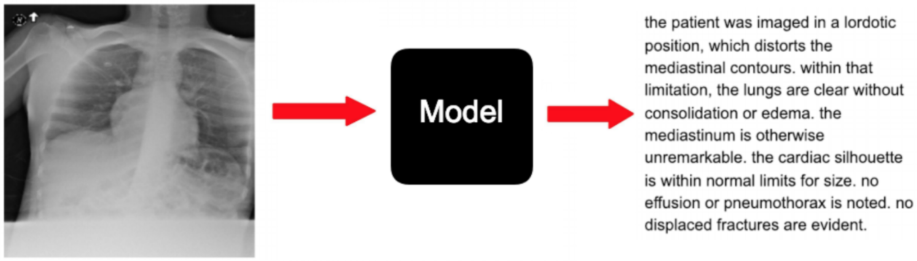
The recent release of many Chest X-Ray datasets has prompted a lot of interest in radiology report generation. Whereas the impression section is meant to be a summarization of the relevant takeaways (given the context, reason for exam, etc.), the findings section is merely meant to convey the factual observations without too much interpretation. Because that would make for a more consistent starting point, automatic report generation has typically been framed as an image captioning task as shown in Figure 2-2, where the machine takes an image of the X-Ray as input and generates the findings section of the radiology report.

2.2 Related Work

2.2.1 Radiology Report Generation

In the last five years, there has been a large increase in the number of Chest X-Ray datasets available for deep learning structured prediction tasks. In 2017, NIH released ChestX-ray8, containing 109,000 frontal X-Ray images where 8 labels were extracted using NLP on non-public radiology reports [299]. In 2019, Stanford and MIT jointly

Figure 2-2: Radiology report generation as an image captioning task.



the patient was imaged in a lordotic position, which distorts the mediastinal contours. within that limitation, the lungs are clear without consolidation or edema. the mediastinum is otherwise unremarkable. the cardiac silhouette is within normal limits for size. no effusion or pneumothorax is noted. no displaced fractures are evident.

released CheXpert and MIMIC-CXR-JPG, respectively, which collectively totalled 450,000 images annotated with the same 14 labels derived from the CheXpert tool [127, 133] The following year, Stanford released CheXphoto, a subset of the CheXpert photos and labels where the images are real-world camera pictures of the image rendered on a screen [127]. In 2021, researchers from Vietnamese hospitals released VinDr-CXR, a 100,000 Chest X-Ray images with 28 labels [201].

It has been more rare to find datasets which make the corresponding radiology reports available to accompany the images. Indiana University released a public Chest X-Ray dataset on Open-I containing 7,000 images and 4,000 reports [67]. The year after MIMIC-CXR-JPG's release, MIT released MIMIC-CXR, which made available the reports for nearly 250,000 radiographic studies [133]. Additionally, PadChest is a public dataset of 160,000 Chest X-Rays along with both structured labels and radiology reports in Spanish [45].

There have been numerous recent attempts to generate reports for X-Ray studies. Gale et al. [91] trained an RNN on templated source data to be able to produce a text description of their structured predictions for pelvic X-Rays. Hsu et al. [122] learned a joint representation for reports and Chest X-Ray images through unsupervised alignment of cross-modal embedding spaces. Zhang et al. [321] use text-based descriptions of patient clinical history, exam technique, prior exam, and radiology findings section to generate a short summary of the radiology impressions section. Wang et al. [300] used a CNN-RNN architecture with attention to generate reports that describe Chest X-Rays. Jing, Xie, and Xing [131] use a CNN to encode the Chest X-Ray image and a hierarchical LSTM with attention to decode it into a written report. Li et al. [162] generated Chest X-Ray reports using reinforcement learning to tune a hierarchical decoder that chooses (for each sentence) whether to use an existing template or to generate a new sentence, optimizing for language fluency metrics.

2.2.2 Deep Learning and Image Captioning

As the deep learning revolution started taking form in 2012–2014, one of the first big tasks for image captioning was from the Microsoft COCO dataset [165]. For a given image, generate a readable, accurate, and linguistically correct caption. This task has received significant attention with the success of *Show and Tell* [293] and its followup *Show, Attend, and Tell* [312]. Due to the leaderboard-style COCO competition, other works quickly emerged showing strong results: Yao et al. [317] used boosting methods, Lu et al. [171] employed adaptive attention, and Rennie et al. [240] introduced reinforcement learning as a method for fine-tuning generated text. Devlin et al. [68] performed surprisingly well using a K-nearest neighbor method. They observed that since most of the true captions were simple, one-sentence scene descriptions, there was significant redundancy in the dataset.

Within the last few years, many have explored the very impressive results of deep learning for text generation. Graves [103] outlined best practices for RNN-based sequence generation. The following year, Sutskever, Vinyals, and Le [278] introduced the *sequence-to-sequence* paradigm for machine translation and beyond. However, Wiseman, Shieber, and Rush [308] demonstrated that while RNN-generated texts are often fluent, they have typically failed to reach human-level quality.

Recent efforts have also begun employing reinforcement learning due to its capability to optimize for indirect target rewards, even when the targets themselves are often non-differentiable. Li et al. [161] used a crafted combination of human heuristics as the reward while Bahdanau et al. [17] incorporated language fluency metrics. They were among the first to apply such techniques to neural language generation, but to date, training with log-likelihood maximization [309] has been the main working horse.

Alternatively, Rajeswar et al. [233] and Fedus, Goodfellow, and Dai [84] have tried using Generative Adversarial Neural Networks (GANs) for text generation. However, Caccia et al. [46] observed problems with training GANs and show that to date, they are unable to beat canonical sequence decoder methods.

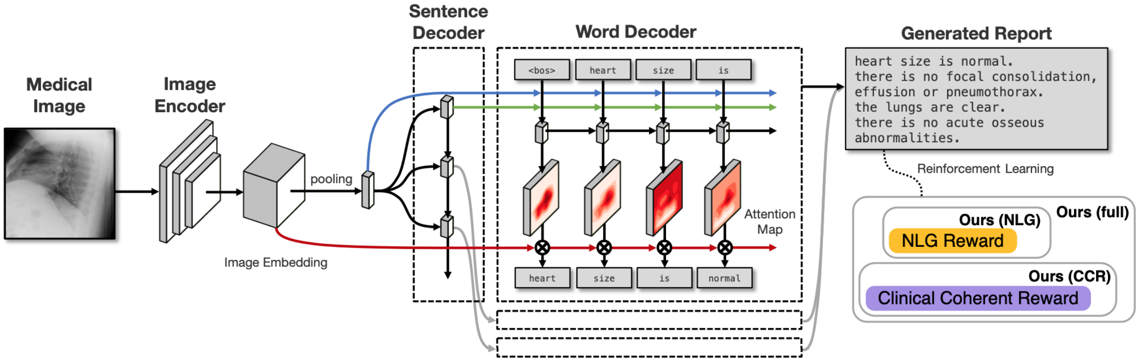
2.3 Model Design

This work focuses on generating a clinically useful radiology report from a Chest X-Ray image. This task has been explored multiple times, but directly transplanting natural language generation techniques onto this task only trains the model to produce reports that *look real* rather than ones that *are clinically correct.* A more immediate focus for the report generation task is thus to produce accurate disease profiles to power downstream tasks such as diagnosis and care providing.

Many traditional image captioning approaches are designed to produce far shorter and less complex pieces of text than radiology reports. Further, these approaches do not capitalize on the highly templated nature of radiology reports. Additionally, generic natural language generation (NLG) methods prioritize descriptive accuracy only as a byproduct of readability, whereas providing an accurate clinical description of the radiograph is the *first* priority of the report. Prior works in this domain have partially addressed these issues, but significant gaps remain towards producing high-quality reports with maximal clinical efficacy.

In this section, I describe our model design in two steps. In subsection 2.3.1, I describe the model architecture. In subsection 2.3.2, I describe a technical improvement to the system motivated by clinical considerations. To summarize briefly the process depicted in Figure 2-3:

1. The model encodes the Chest X-Ray image as a fixed-dimensional embedding.
2. The image embedding is decoded into a sequence of unconstrained "topic vectors" by a sentence-level LSTM.
3. Each "topic vector" is decoded into a sequence of words by a word-level LSTM, using attention over the original image.
4. The model is optimized using reinforcement learning in order to maximize both a traditional natural language generation score (CIDEr [291]) and a clinical accuracy score (CheXpert [127]).



heart size is normal.
there is no focal consolidation,
effusion or pneumothorax.
the lungs are clear.
there is no acute osseous
abnormalities.

Figure 2-3: The model for clinically-accurate report generation.

2.3.1 CNN-RNN-RNN Model Architecture

As illustrated in Figure 2-3, we aim to generate a report as a sequence of sentences $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_M)$, where M is the number of sentences in a report. Each sentence consists of a sequence of words $\mathbf{z}_i = (z_{i1}, ..., z_{iN_i})$ with words from a vocabulary $v_{ij} \in \mathbb{V}$, where N_i is the number of words in sentence i.

The image is fed through the *image encoder CNN* to obtain a visual feature map. The feature is then taken by the *sentence decoder RNN* to recurrently generate vectors that represent the topic for each sentence. With the visual feature map and the topic vector, a *word decoder RNN* tries to generate a sequence of words and attention maps of the visual features. This hierarchical approach is in line with Krause et al. [155] where they generate descriptive paragraphs for an image.

Image encoder CNN The input image I is passed through a CNN head to obtain the last layer before global pooling, and the feature is then projected to an embedding of dimensionality d, which is identical to the word embedding dimension. The resulting map $\mathbf{V} = \{\mathbf{v}_k\}_{k=1}^{K}$ of spatial image features will be descriptive features for different spatial locations of an image. A mean visual feature is obtained by averaging all local visual features

$$\bar{\mathbf{v}} = \frac{1}{K} \sum_k \mathbf{v}_k \tag{2.1}$$

Sentence decoder RNN Given the mean visual feature $\bar{\mathbf{v}}$, we adopt Long Short-Term Memory (LSTM) and model the hidden state as

$$\mathbf{h}_i, \mathbf{m}_i = \mathrm{LSTM}(\bar{\mathbf{v}}; \mathbf{h}_{i-1}, \mathbf{m}_{i-1}), \tag{2.2}$$

where $\mathbf{h}_{i-1}$ and $\mathbf{m}_{i-1}$ are the hidden state vector and the memory vector for the previous sentence $(i-1)$ respectively. From the hidden state $\mathbf{h}_i$, we further generate two components, namely the topic vector $\boldsymbol{\tau}_i$ and the stop signal u_i for the sentence, as

$$\boldsymbol{\tau}_i = \mathrm{ReLU}(\mathbf{W}_\tau^\top \mathbf{h}_i + \mathbf{b}_\tau) \tag{2.3}$$

$$u_i = \sigma(\mathbf{w}_u^\top \mathbf{h}_i + b_u) \tag{2.4}$$

where $\mathbf{W}$'s and $\mathbf{b}$'s are trainable parameters, and σ is the sigmoid function. The stop signal acts as as the end-of-sentence token. When $u > 0.5$, it indicates the sentence decoder RNN should stop generating the next sentence.

Word decoder RNN After we decode the sentence topics, we can start to decode the words given the topic vector $\boldsymbol{\tau}_i$. For simplicity, we drop the subscript i as this process applies to each sentence. We adopted the visual sentinel [171] that modulates the feature map $\mathbf{V}$ with a sentinel vector. This formulation enables the model to look at different parts on the image while having the option of "looking away" at a sentinel vector. The hidden states and outputs are again modeled with LSTM, generating the posterior probability $\mathbf{p}_j$ over the vocabulary with (1) the mean visual feature $\bar{\mathbf{v}}$, (2) the topic vector $\boldsymbol{\tau}$, and (3) the embedding of the previously generated word $\mathbf{e}_{j-1} = \mathbf{E}_{z_{j-1}}$, where $\mathbf{E} \in \mathbb{R}^{d \times |\mathbb{V}|}$ is the trainable word embedding matrix. At training time, the next word is sampled from the probability $z_j \sim p(z \mid \cdot) = (\mathbf{p}_j)_z$, or the z-th element of $\mathbf{p}_j$.

We calculate the sentinel vector $\mathbf{s}_j$, the attention over the K regions of the images and the sentinel gate $\hat{\boldsymbol{\alpha}}_j$, the mixture context vector $\hat{\mathbf{c}}_j$, and the probability $\mathbf{p}_j$ over the vocabulary as

$$\begin{aligned}
\mathbf{h}_j, \mathbf{s}_j, \mathbf{m}_j &= \text{LSTM}\left([\bar{\mathbf{v}}, \boldsymbol{\tau}, \mathbf{e}_{j-1}]; \mathbf{h}_{j-1}, \mathbf{m}_{j-1}\right) && (2.5)\\
\boldsymbol{\alpha}_j &= \mathbf{w}_\alpha^\top \tanh\left(\left[\mathbf{W}_{v\alpha}^\top \mathbf{V}, \mathbf{W}_{s\alpha}^\top \mathbf{s}_j\right] + \left(\mathbf{W}_{h\alpha}^\top \mathbf{h}_j\right)\mathbf{1}^\top\right) && (2.6)\\
\hat{\boldsymbol{\alpha}}_j &= \text{softmax}\left(\boldsymbol{\alpha}_j\right) && (2.7)\\
\hat{\mathbf{c}}_j &= [\mathbf{V}, \mathbf{s}_j]\hat{\boldsymbol{\alpha}}_j && (2.8)\\
\mathbf{p}_j &= \text{softmax}\left(\mathbf{W}_p^\top(\hat{\mathbf{c}}_j + \mathbf{h}_j)\right), && (2.9)
\end{aligned}$$

where $\mathbf{h}_{j-1}$ and $\mathbf{m}_{j-1}$ again are the hidden state vector and the memory vector for the previous step, $\mathbf{W}$'s are weights to be learned. $(\cdot, \cdot)$ denotes matrix concatenation, and $\mathbf{1}$ denotes a vector of all one's. Note that this hierarchical encoder-decoder CNN-RNN-RNN architecture is fully differentiable.

We observed our model to sometimes repeat the findings multiple times. We apply post-hoc processing where we remove exact duplicate sentences in the generated reports. This proves to improve the readability but interestingly slightly degrades NLG metrics.

Full implementation details, including layer sizes, training details, etc., are presented in the Appendix A.

2.3.2 Clinical Coherence

One major downside with the approach outlined so far is that traditional objective functions would prioritize generating output that looks good but fails to distinguish between reports that are factually correct (e.g. “pleural effusion is present”) and factually incorrect (e.g. “pleural effusion is not present”). Negative judgments on diseases are critical components of the reports; the ordering physician cares much more about getting the correct information than they do about the readability.

To measure the clinical correctness of a given report, we use the CheXpert sentence labeler, which is a rule-based system that extracts 14 categories[1] of key findings from

[1] 12 types of thoracic diseases or X-Ray related diagnoses, the presence of support devices, and “no finding”.

a report. By feeding both the generated text and the associated image's reference report to CheXpert, we are able to identify which findings were correctly identified, as well as false positive and false negative findings. Although the labeler does require a baseline level of grammaticality, this evaluation largely ignores readability or marginal increases in grammaticality.

Because naively optimizing with cross entropy or traditional language models may misguide the model to mention only the disease names as opposed to correctly positively/negatively describe the disease states, we incorporate CheXpert directly into the model's objective function. Because CheXpert is not differentiable with respect to the model parameters, we demonstrate how reinforcement learning can propagate the clinical signal to the model. We propose training the model with this CheXpert-derived "Clinically Coherent Reward" (CCR).

We consider the case of self-critical sequence training (SCST) [240] which utilizes the REINFORCE [305] algorithm, and minimize the negative expected reward as a function of the network parameters θ, as

$$\mathcal{L}_{\text{NLG}}(\theta) = -\mathbb{E}_{(u,\mathbf{Z})\sim p_\theta(u,\mathbf{Z})}[r_{\text{NLG}}(\mathbf{Z},\mathbf{Z}^*) - r_{\text{NLG}}(\mathbf{Z}^g,\mathbf{Z}^*)], \tag{2.10}$$

where p_θ is the distribution over output spaces, r_{NLG} is a metric evaluation function acting as a reward function that takes a sampled report $\mathbf{Z}$ and a ground truth report $\mathbf{Z}^*$. The baseline in SCST has been replaced with the reward obtained with testing time greedily decoded report $\mathbf{Z}^g$.

For each label type t in CheXpert, there are four possible outcomes for the labeling: (1) positive, (2) negative, (3) uncertain, or (4) absent mention; or, $l_t(\mathbf{Z}) \in \{\mathsf{p},\mathsf{n},\mathsf{u},\mathsf{a}\}$. This outcome can be used to model the positive/negative disease state $s_t \in \{+,-\}$ as $s_t \sim p_{s|l}(\cdot|l_t(\mathbf{Z}))$, the value of which will be discussed further later. CCR is then defined, dropping the subscripts for distribution for convenience, as

$$r_{\text{CCR}}(\mathbf{Z},\mathbf{Z}^*) = \sum_t r_{\text{CCR},t}(\mathbf{Z},\mathbf{Z}^*) \equiv \sum_t \sum_{s\in\{+,-\}} p(s|l_t(\mathbf{Z}))\cdot p(s|l_t(\mathbf{Z}^*)), \tag{2.11}$$

aiming to maximize the correlation of distribution over disease states between the generated text $\mathbf{Z}$ and the ground truth text $\mathbf{Z}^*$. Unfortunately, as the true diagnostic state s of novel reports is unknown, we need to make several assumptions regarding the performance of the rule based labeler, allowing us to infer the necessary conditional probabilities $p(s|l)$.

To motivate these assumptions, first note that these diseases are universally rare, or, $p(+) \ll p(-)$. Presuming the rule based labeler has any discriminative power, we can thus conclude that if the labeler assigns a negative or an absent label (l^- is one of $\{\mathsf{n}, \mathsf{a}\}$), $p(+|l^-) < p(+) \ll p(-) < p(-|l^-)$. For sufficiently rare conditions, a reasonable assumption and simplification is to therefore take $p(+|l^-) \approx 0$ and $p(-|l^-) \approx 1$. We further assume that the rule based labeler has a very high precision, and thus $p(+|\mathsf{p}) \approx 1$. However, given an uncertain mention u, the desired output probabilities are difficult to assess. As such, we define a reward-specific hyperparameter $\beta_{\mathsf{u}} \equiv p(+|\mathsf{u})$, which in this work we take to be 0.5. All of these assumptions could be easily adjusted, but they perform well for us here.

We also wish to use a baseline for the reward r_{CCR}. Instead of using a single exponential moving average (EMA) over the total reward, we apply EMA separately to each term as

$$\mathcal{L}_{CCR}(\theta) = -\mathbb{E}_{(u,\mathbf{Z})\sim p_\theta(u,\mathbf{Z})}\left[\sum_t r_{CCR,t}(\mathbf{Z}, \mathbf{Z}^*) - \bar{r}_{CCR,t}\right], \tag{2.12}$$

where $\bar{r}_{CCR,t}$ is an EMA over $r_{CCR,t}$ updated as

$$\bar{r}_{CCR,t} \leftarrow \gamma \bar{r}_{CCR,t} + (1-\gamma) r_{CCR,t}(\mathbf{Z}, \mathbf{Z}^*) \tag{2.13}$$

We wish to pursue both semantic alignment and clinical coherence with the ground truth report, and thus we combine the reinforcement learning rewards for CheXpert (clinical correctness, i.e. “CCR”) and CIDEr (readability, i.e. “NLG”) in a weighted fashion. Specifically, $\mathcal{L}(\theta) = \mathcal{L}_{\text{NLG}}(\theta) + \lambda \mathcal{L}_{CCR}(\theta)$, where λ controls the relative importance.

Hence the derivative of the combined loss with respect to θ is thus

$$\nabla_\theta \mathcal{L}(\theta) = -\mathbb{E}_{(u,\mathbf{Z})\sim p_\theta(u,\mathbf{Z})}\left[[r_{\mathrm{NLG}}(\mathbf{Z},\mathbf{Z}^*) + \lambda r_{\mathrm{CCR}}(\mathbf{Z},\mathbf{Z}^*)]\nabla_\theta \sum_i \left(\log u_i + \sum_j \log\left(\mathbf{p}_{ij}\right)_{z_{ij}}\right)\right], \tag{2.14}$$

where $\mathbf{p}_{ij}$ is the probability over vocabulary. We can approximate the above gradient with Monte-Carlo samples from p_θ and average gradients across training examples in the batch.

2.4 Methodology

2.4.1 Data

In this work, we use two Chest X-Ray datasets: MIMIC-CXR [133] and Open-I [67].

MIMIC-CXR is the largest radiology dataset to date and consists of $473,057$ Chest X-Ray images and $206,563$ reports from $63,478$ patients[2]. Among these images, $240,780$ are of anteroposterior (AP), $101,379$ are of posteroanterior (PA), and $116,023$ are of lateral (LL) views. Furthermore, we eliminate duplicated radiograph images with adjusted brightness level or contrast as they are commonly produced for clinical needs, after which we are left with $327,281$ images and $141,783$ reports. The radiological reports are parsed into sections, among which we extract the *findings* section. We then apply tokenization and keep tokens with at least 5 occurrences in the corpus, resulting in $5,571$ tokens in total.

Open-I is a public radiography dataset collected by Indiana University with $7,471$ Chest X-Ray images and $3,955$ reports. The reports are in XML format and include pre-parsed sections. We then exclude the entries without the *findings* section and are left with $6,471$ images and $3,336$ reports. Tokenization is done similarly, but due to the relatively small size of the corpus, we keep tokens with 3 or more occurrences, ending up with 948 tokens.

[2] This work used an alpha version of MIMIC-CXR instead of the publicly released version where the images are more standardized and split into official train/test sets.

Both datasets are partitioned by patients into a train/validation/test ratio of 7/1/2 so that there is no patient overlap between sets. Words that are excluded were replaced by an "unknown" token, and the word embeddings are pretrained separately for each dataset.

2.4.2 Models

We compare our methods with state-of-the-art image captioning and medical report generation models as well as some simple baseline models:

- *1-NN*, in which we query in the image embedding space for the closest neighbor in the train set. The corresponding report of the neighbor is used as the output for this test image;
- *Show and Tell* (S&T) [293];
- *Show, Attend, and Tell* (SA&T) [312]; and
- *TieNet* [300].

To allow comparable results in all models, we slightly modify previous models to also accept the view position embedding which encodes AP/PA/LL as a one-hot vector to utilize the extra information available at image acquisition. This includes *Show and Tell*, *Show, Attend, and Tell*, and our re-implementation of *TieNet*, which is detailed in Appendix A.2 because the authors did not release their code.

Additionally, we train a baseline LSTM which generates free text without conditioning on input radiograph images, which we denote as *Noise-RNN*. The purpose of this model is to serve as a sanity check by contextualizing how easy/hard this task is (e.g., report complexity, report variance, sensitivity of evaluation metrics, etc.).

To evaluate our model, we perform several ablation studies to inspect the contribution of various components of our model. In particular, we assess

- Ours (NLG): Use r_{NLG} only for reinforced learning, as often is the case with the prior state-of-the-art.
- Ours (CCR): Use r_{CCR} only and do not care about aligning the natural language metrics.
- Ours (full): Considers both rewards.

All recurrent models, including prior works and our models, use beam search with a beam size of 4.

2.4.3 Evaluation

To compare with other models including prior state-of-the-art and baselines, we adopt several different metrics that focus on different aspects ranging from a natural language perspective to clinical adequacy.

There are a number of evaluation metrics for text generation approaches, such as BLEU [221], CIDEr [291], METEOR [19], ROUGE [163], and SPICE [9]. Even in the general domain, these metrics have known shortcomings related to their propensity for favoring surface-level similarity in texts [32] and exhibiting weak correlations with human judgment [48]. Nonetheless, there are likely to be shortcomings with any automatic method, and researchers need some standardized metrics. Wang et al. [300] employ BLEU, METEOR, and ROUGE; Xu et al. [311] evaluate with BLEU-4, METEOR, and CIDEr; Gale et al. [91] use BLEU; the COCO leaderboard uses BLEU, METEOR, ROUGE and CIDEr. In this work, we use BLEU and ROUGE — two of the oldest and most widespread metrics — and CIDEr — a more recently developed metric which has intitutive design and stronger demonstrated correlation with annotations [291]

One concern with such statistical measures is that with a limited scope from the n-grams (n up to 4) we are unable to capture disease states, as previously discussed in Section 2.3.2. As such, we also include CheXpert-derived scores as metrics. Specifically, we compare the generated text to the reference text to compute the true positives, false positives, and false negatives. With this, we then derive the accuracy and precision of the generated text.

	Model	**Natural Language**						**Clinical**
		CIDEr	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Accuracy
MIMIC-CXR	*Major Class*	-	-	-	-	-	-	0.828
	Noise-RNN	0.716	0.272	0.269	0.172	0.113	0.074	0.803
	1-NN	0.755	0.244	0.305	0.171	0.098	0.057	0.818
	S&T	0.886	0.300	0.307	0.201	0.137	0.093	0.837
	SA&T	0.967	0.288	0.318	0.205	0.137	0.093	0.849
	TieNet	1.004	0.296	0.332	0.212	0.142	0.095	0.848
	Ours (NLG)	**1.153**	**0.307**	**0.352**	**0.223**	**0.153**	**0.104**	0.834
	Ours (CCR)	0.956	0.284	0.294	0.190	0.134	0.094	**0.868**
	Ours (full)	1.046	0.306	0.313	0.206	0.146	0.103	0.867
Open-I	*Major Class*	-	-	-	-	-	-	0.911
	Noise-RNN	0.747	0.291	0.233	0.130	0.087	0.061	0.914
	1-NN	0.728	0.201	0.232	0.116	0.051	0.018	0.911
	S&T	0.926	0.306	0.265	0.157	0.105	0.073	0.915
	SA&T	1.276	0.313	0.328	0.195	0.123	0.080	0.908
	TieNet	1.334	0.311	0.330	0.194	0.124	0.081	0.902
	Ours (NLG)	**1.490**	**0.359**	**0.369**	**0.246**	**0.171**	**0.115**	0.916
	Ours (CCR)	0.707	0.244	0.162	0.084	0.055	0.036	0.917
	Ours (full)	1.424	0.354	0.359	0.237	0.164	0.113	**0.918**

Table 2.1: **Automatic Evaluation Scores.** The table is divided into natural language metrics and clinical finding accuracy scores. BLEU-n counts up n-gram for evaluation, and accuracy is the averaged macro accuracy across all clinical findings. *Major class* always predicts negative findings.

2.5 Results

2.5.1 Quantitative Results

Overall Model Performances Table 2.1 shows a comprehensive assessment for baseline models, prior works, and variants of our model, evaluated using BLEU, ROUGE, CIDEr, and CheXpert accuracy.

The Noise-RNN baseline establishes the floor of "completely unremarkable performance" for each metric. This can be very useful for trying to suss out how impressive a given model's performance is. For instance, we can see that the 1-NN model was virtually on par with Noise-RNN, suggesting that unlike COCO (with sentence complexity "A car parked in front of a building"), radiology reports are much more complex and far less interchangeable.

We see that, as expected, Show and Tell performs marginally worse than Show, Attend, and Tell, which in turn performs very similarly to TieNet. Additionally, our full model outperforms those prior works.

The ablation study demonstrates that when we exclusively optimise for CIDEr, our model does the best on NLG metrics, at the expense of clinical accuracy. Similarly, when we exclusively optimize for CheXpert correctness, our model does the best on CheXpert metrics, at great expense to performance on natural language metrics. The full model is able to maintain the strong performance on clinical correctness as the CCR-only model while still enjoying large gains in natural language metrics.

Clinical Efficacy In Table 2.2 we examine the clinical correctness performance more closely. Because the labeling process generates discrete binary labels as opposed to predicting continuous probabilities, we are unable to obtain discriminative metrics such as the Area Under the Receiver Operator Characteristic (AUROC). Recognizing that accuracy is not a good metric and that there may be noteworthy trends across the 14 labels, we report per-label and aggregate CheXpert precision.

Our model achieves the best aggregate precision; the full version outperforms all other baselines and prior work, and when we optimize for CCR exclusively, we achieve

	Label	Count	1-NN	S&T	SA&T	TieNet	Ours (NLG)	Ours (CCR)	Ours (full)
MIMIC-CXR	Total	69,031	-	-	-	-	-	-	-
	Support Devices	22,227	0.534	0.823	0.847	0.827	0.794	0.849	**0.880**
	Airspace Opacity	21,972	0.432	0.607	0.592	0.571	0.453	**0.640**	0.460
	Cardiomegaly	19,065	0.440	0.535	0.438	0.464	0.000	0.678	**0.704**
	Atelectasis	16,161	0.374	0.490	0.496	0.470	0.385	0.476	**0.521**
	No Finding	15,677	0.432	0.299	0.349	0.339	0.339	**0.491**	0.405
	Pleural Effusion	15,283	0.534	0.550	0.545	**0.735**	0.487	0.683	0.689
	Edema	6,594	0.265	0.331	0.244	**0.405**	0.266	0.280	0.000
	Enlarged Cardiomediastinum	6,064	0.123	0.134	0.163	0.179	0.180	**0.202**	0.167
	Pneumonia	3,068	0.065	0.106	0.091	0.082	0.075	0.000	**0.400**
	Pneumothorax	2,636	0.079	0.034	0.095	0.081	0.081	0.039	**0.098**
	Fracture	2,617	**0.059**	0.000	0.000	0.000	0.000	0.000	0.000
	Lung Lesion	2,447	0.064	**0.333**	0.223	0.000	0.000	0.000	0.000
	Consolidation	2,384	0.076	0.013	**0.180**	0.151	0.089	0.037	0.000
	Pleural Other	1,285	**0.039**	0.000	0.103	0.000	0.000	0.000	0.000
	Precision (macro)		0.253	0.304	**0.312**	0.307	0.225	**0.313**	0.309
	Precision (micro)		0.383	0.414	0.430	0.473	0.419	**0.634**	0.586

Table 2.2: **Clinical Finding Scores**. The precision scores for each of the labels are listed and aggregated into the overall precision scores. Macro denotes averaging the numbers in the table directly and micro accounts for class prevalence. Rows are ordered by class prevelance.

even higher performance. We notice that the higher-performing models (full, CCR, and TieNet to some extent) perform very well on highly prevelant tasks, and have precisions of 0.0 on very rare labels. The poorly-performing models (1-NN, S&T to some extent) perform marginally better on the low-prevelance labels but notably worse on the common labels.

2.5.2 Qualitative Results

Evaluation of Generated Reports Table 2.3 demonstrates the qualitative results of our full model. In general, our models are able to generate descriptions that align with the logical flow of reports written by radiologists, which start from general information (such as views, previous comparison), positive, then negative findings, with the order of lung, heart, pleura, and others. TieNet also generates report descriptions with such logical flow but in slightly different orders. For the negative findings cases, both our model and TieNet do well on generating reasonable descriptions without significant errors. Regarding the cases with positive findings, TieNet and our full

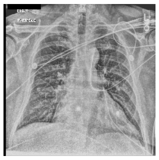
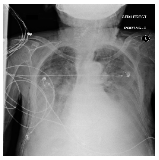
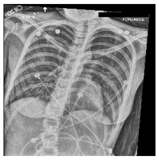
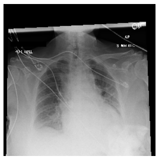
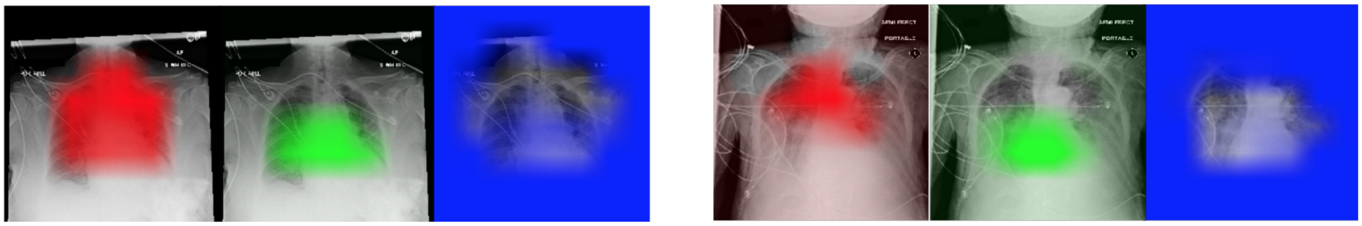
	Ground Truth	TieNet	Ours (full)
	cardiomegaly is moderate. bibasilar atelectasis is mild. there is no pneumothorax. a lower cervical spinal fusion is partially visualized. healed right rib fractures are incidentally noted.	ap portable upright view of the chest. there is no focal consolidation, effusion, or pneumothorax. the cardiomediastinal silhouette is normal. imaged osseous structures are intact.	pa and lateral views of the chest. there is mild enlargement of the cardiac silhouette. there is no pleural effusion or pneumothorax. there is no acute osseous abnormalities.
	as compared to the previous radiograph, the monitoring and support devices are unchanged. unchanged bilateral pleural effusions, with a tendency to increase, and resultant areas of atelectasis. the air collection in the bilateral soft tissues is slightly decreased. unchanged right picc line. no definite evidence of pneumothorax.	as compared to the previous radiograph, the patient has received a nasogastric tube. the course of the tube is unremarkable, the tip of the tube projects over the middle parts of the stomach. there is no evidence of complication, notably no pneumothorax. the other monitoring and support devices are constant. constant appearance of the cardiac silhouette and of the lung parenchyma.	as compared to the previous radiograph, there is no relevant change. tracheostomy tube is in place. there is a layering pleural effusions. NAME bilateral pleural effusion and compressive atelectasis at the right base. there is no pneumothorax.
	the course of the dobbhoff feeding catheter is unremarkable, and its tip is seen projecting over the stomach. there is no evidence of complications, specifically no pneumothorax. as compared to the prior radiograph dated DATE, there has been no other significant interval change.	ap portable upright view of the chest. overlying ekg leads are present. there is no focal consolidation, effusion, or pneumothorax. the cardiomediastinal silhouette is normal. imaged osseous structures are intact.	as compared to the previous radiograph, there is no relevant change. the endotracheal tube terminates approximately 3 cm above the NAME. the endotracheal tube extends into the stomach. there is no evidence of complications, notably no pneumothorax. there is no pleural effusion or pneumothorax.
	interval placement of a left basilar pigtail chest tube with improving aeration in the left mid to lower lung and near complete resolution of the pleural effusion. there are residual patchy opacities within the left mid and lower lung as well as at the right base favoring resolving atelectasis. no pneumothorax. heart remains stably enlarged. mediastinal contours are stably widened, although this NAME be related to portable technique and positioning. this can be better evaluated on followup imaging. no pulmonary edema.	as compared to the previous radiograph, the patient has been extubated. the nasogastric tube is in unchanged position. the lung volumes remain low. moderate cardiomegaly with minimal fluid overload but no overt pulmonary edema. no larger pleural effusions. no pneumonia.	ap upright and lateral views of the chest. there is moderate cardiomegaly. there is no pleural effusion or pneumothorax. there is no acute osseous abnormalities.

Table 2.3: **Sample images along with ground truth and generated reports**. Note that upper case tokens are results of anonymization.

model both cannot identify all radiological findings. Our full model is able to identify the major finding in each demonstrated case. For example, cardiomegaly in the first case, pleural effusion and atelectasis in the second case.

A formerly practicing clinician co-author reviewed a larger subset of our generated reports manually. They drew several conclusions. First, our full model tends to generate sentences related to pleural effusion, atelectasis, and cardiomegaly correctly—which is aligned with the clinical finding scores in Table 2.2. TieNet instead misses some positive findings in such cases. Second, there are significant issues in *all* generated reports, regardless of the source model, which include the description of supportive lines and tubes, as well as lung lesions. For example, TieNet is prone to generate nasogastric tube mentions while our model tends to mention tracheostomy or endotracheal tube, and yet both models have difficulty identifying some specific lines such as chest tube or PICC line. Similarly, both systems do not generate the sentence with positive lung parenchymal findings correctly.

Learning Meaningful Attention Maps Attention maps have been a useful tool in visualizing what a neural network is attending to, as demonstrated by Rajpurkar et al. [237]. Figure 2-4 shows the intermediate attention maps for each word when it is being generated. As we can observe, the model is able to roughly capture the location of the indicated disease or parts, but we also find, interestingly, that the attention map tends to be the complement of the actual region of interest when the disease keywords follow a negation cue word. This behavior has not been widely discussed before, partially because attention maps for negations are not the primary focus of typical image captioning tasks, and most attention mechanisms employed in a clinical context were on classification tasks where they also do not specifically focus on negations.



ap upright and lateral views of the chest. there is moderate cardiomegaly. there is no pleural effusion or pneumothorax. there is no acute osseous abnormalities.

(a)

as compared to the previous radiograph, there is no relevant change. tracheostomy tube is in place. there is a layering pleural effusions. NAME bilateral pleural effusion and compressive atelectasis at the right base. there is no pneumothorax.

(b)

Figure 2-4: **Visualization of the generated report and image attention maps.** Different words are underlined with its corresponding attention map shown in the same color. Best viewed in color.

2.6 Project "Evidence"

In this work, we develop a Chest X-Ray radiology report generation system which hierarchically generates topics from images, then words from topics. The final system is also optimized with reinforcement learning for both readability (via CIDEr) and clinical correctness (via the novel Clinically Coherent Reward). Our system outperforms a variety of compelling baseline methods across readability and clinical efficacy metrics on both MIMIC-CXR and Open-I datasets.

The goals of this section are two-fold. First, I apply tools from Effective Altruism and Data Feminism to this project as a case study to demonstrate how one can think through the choice and impact of potential research directions they are considering. Second, I examine the impact this specific project has had in the three years since it was published.

2.6.1 Evaluating a Choice of Project

Effective Altruism

As discussed in Section 1.5.1, many effective altruists adopt the "ITN" Framework (Importance, Tractability, and Neglected-ness) to measure the expected impact of a given organization or program.

- Importance: How useful would it be to the world if this problem was solved?
- Tractability: How much of a difference can one make towards solving this problem?
- Neglected-ness: How over-saturated is the space of solutions?

The EA organization 80,000 Hours elaborates further[3] on how they implement these values. The assessments do not occur in a vacuum, but are meant to help compare multiple choices (e.g., where to donate $1000, what kind of career to pursue). Although some of the concepts will need translation, these tools can be a helpful starting point in assessing the impact of this project's scope.

The importance of the effort, of course, depends on what kind of effort is done. The expected value of the project is a function of both 1) how meaningful would the project be if it is successful; and 2) how likely is the project to be successful. 80,000 Hours estimated[4] "Positively shaping the development of artificial intelligence" to be one of the highest priority areas because even though investing in research is less likely to pay off in the short-term than investing in scaling up a known intervention, they estimated the upside of AI safety is very, very high. They also considered investing in short-term interventions, and found that the most effective approaches are ones that deliver life-saving services to residents of developing countries[5] (e.g., malaria bed nets, sustained TB treatment with antibiotics, anti-retroviral drugs for HIV, etc). Therefore, the scoping of a project's ambitions is done in conjunction with the other two factors of tractability and neglected-ness to choose a project that has the best expected value for one's risk tolerance.

Importance. Within the US, there are opportunities for technology to speed up care, just as telehealth has done, and to improve the workflow of radiologists. Some companies like viz.ai have been able to obtain FDA approval to market AI for radiology products which fit into the workflow for the care team and for which hospitals have paid $1,000 per patient [209]. However, in general, the healthcare sector

[3] https://80000hours.org/articles/problem-framework

[4] https://80000hours.org/problem-profiles/positively-shaping-artificial-intelligence

[5] https://80000hours.org/problem-profiles/health-in-poor-countries

is slow to adopt new technological advances for many reasons, including ambiguity related to federal-stare law interactions, reimbursements, licensure, efficacy, ease of use, data sharing and privacy, fear of lawsuits, and more [159]. Typically, effective altruists have found that interventions in the developing world are able to have orders of magnitude more impact because of greater needs and fewer barriers [174].

Radiology report generation can be a very important area to work on. In the United States, radiologists' share of the overall physician workforce declined by 8.8% over the last 20 years [246]. A patient safety study at the US Department of Veterans Affairs found that unread radiology exams — caused by unfinished dictations, technologist errors, and inefficient radiologist work tools — resulted in delays to critical patient care [20]. Under-resourced areas, including rural regions, developing countries — and the intersection thereof — face an even more dire situation. In 2020, Rwanda (population 13 million) only had 15 radiologists, and nearly 75% of them were based in Kigali, which contains less than 10% of the country's population [249, 247].

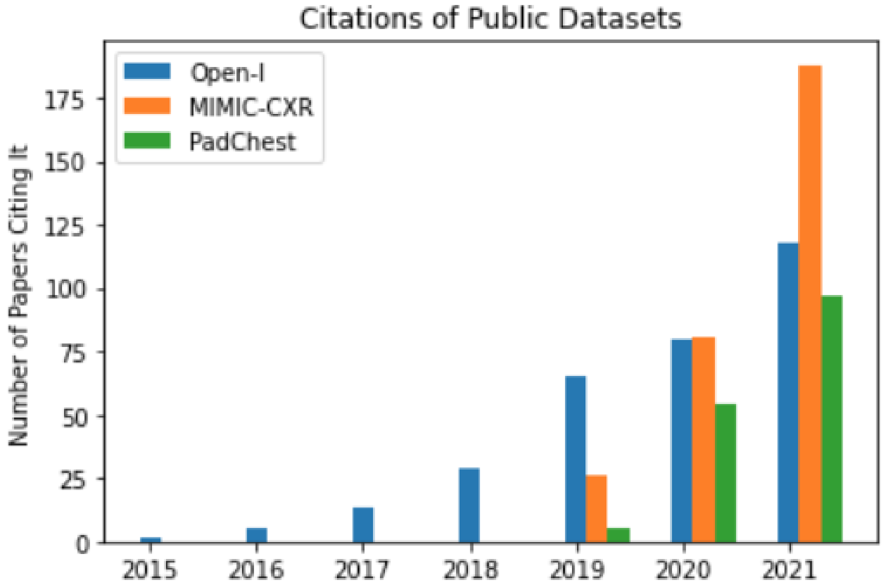
Although EAs try to use precise or pseudo-precise measurements[6] like Quality-adjusted Life Years (QALYs) to quantify just how impactful something could be, I don't think such estimates are feasible or appropriate to apply to the selection of a research topic.

Neglected-ness. When trying to estimate the expected impact of an intervention, effective altruists attempt to forecast the potential outcomes (colloquially known as the "counterfactuals"). The question of interest is "How much better would the world be if I worked on this problem as opposed to if I didn't work on this problem?" Therefore, it is a question of *marginal* value, meaning that ignoring low-hanging fruit or startup costs, how much more value would come from one more unit of input (e.g., one more hour worked, one more dollar donated, etc.). Areas which others have not worked on as much are less likely to have their low-hanging fruit picked already. For this reason, 80,000 Hours uses a "crowdedness score" based on how much funding or full-time staff different causes have.[7]

[6] https://80000hours.org/articles/problem-framework/#definition

[7] https://80000hours.org/articles/problem-framework/#how-to-assess-it-2

Figure 2-5: Number of papers citing each public diagnostic captioning dataset's paper per year.



One way to approximately quantify crowdedness of this task's space is to look at usage of the main datasets over time. Pavlopoulos et al. [227] identified five publicly available datasets for diagnostic captioning: IU X-RAY Open-I [67], MIMIC-CXR [133], PadChest [45], PEIR GROSS [137], ICLEFCAPTION [93]. However, they conclude that PEIR GROSS and ICLEFCAPTION suffer from "severe shortcomings" both in size (less than a few hundred images) and quality (contain photographs and captions from the figures of scientific articles, instead of real diagnostic medical images). CheXpert [127] and ChestX-Ray8 [299] both release large Chest X-Ray datasets but without any public reports, thus making them ill-suited as proxies for interest in report generation. Therefore, we focus on Open-I, MIMIC-CXR, and PadChest.

Figure 2-5 shows how many times each of these three datasets' papers have been cited each year since being published. The release of Chest X-Ray datasets with non-public reports in 2018 (Chest X-Ray8 [299]) and January 2019 (CheXpert [127] and MIMIC-CXR-JPG [133]) created a lot of interest in X-Ray classification and segmentation tasks. In early 2019, there was much research attention energy dedicated to radiology report generation.

Tractability. Different endeavours have different risk profiles for likelihood to succeed. Machine learning for healthcare is an applied ML field, whose ultimate goal is usually to use tools in practice to help people. One of the most common kinds of project researchers work on is to build proof-of-concept models on observational/retrospective data to see whether a task is worth investing more time and effort into [232, 156, 112]. Researchers also develop tools that are useful to the community [127, 4]. If a task/model shows promise for eventual deployment, more research effort is invested into evaluating it more thoroughly, such as studying more realistic framings of how the model would fit into a workflow [90], scrutinizing models for bias [257], comparing against human judgment instead of automatic metrics [34], and investigating how performance holds under dataset shift [200]. Tasks and models that have sufficiently matured may eventually get tested in real-life hospitals and clinics, either through research [256, 315] or through for-profit products [23].

This project is an example of early stage problem exploration. Image-based ML has proven more successful than other forms such as language, though innovations in deep learning continued to improve NLP in the last few years [16, 241, 155]. Many research projects take general domain ML models and adapt them to healthcare settings (e.g., predicting on healthcare data [4]; changing the structure of the model to account for missing data [55], disease progression dynamics [298], etc.). Examples of early-field contributions can make include building tools other researchers benefit from and helping shape the task definition and initial approaches.

This research effort was the first to publish radiology report generation on MIMIC-CXR because it had access to an alpha version of that dataset pre-release. This opportunity allowed us to shape some of the approaches that other papers followed, which is elaborated on further in Section 2.6.2.

Data Feminism

The lens of Data Feminism offers another perspective in the formulation of this task and its potential consequences.

Examine Power. The first principle of Data Feminism is to examine power.

Within the US, radiologists are currently still the authority for reading X-Rays; however, technological advancements have been chipping away at this power [294]. As machine learning augments human ability to understand the X-Ray, this “de-skills” the task. However, that direction is not pre-ordained. An alternative path forward is that AI tools are built for the radiologists, themselves, in order to help them do their job in a faster and more consistent way. Electronic Health Record (EHR) configurations have codified and imposed workflows onto healthcare workers [294]. If this new technology were to be deployed, it would further change the workflow, resulting in some winners and some losers. One can consider the different interests of the involved stakeholders:

- Radiologist: help them read the images (e.g. faster, more consistently, safer, etc.).
- Ordering doctor: answer the question they are studying about what is wrong with the patient.
- Patient: deliver better, safer care equitably.
- Hospital: standardize practice (e.g. efficiency, enforce compliance with internal procedures) and more easily bill the insurer.
- Insurer: check justification for why patient was billed for certain procedures.

We have already begun to see some of the basic, structured radiology AI predictions be deployed [150, 315]. As more tools are brought to practice, there will likely be power struggles among the stakeholders.

Challenge Power. Beyond the US, radiology report generation has the potential to play a large role in healthcare for developing countries. Rwanda barely has 1 radiologist per million people (whereas US has 100 radiologists per million), and most of them are clustered in a single city [249, 247]. The Rwandan startup Insightiv Technologies addresses this issue using technology; it developed a platform which uses both teleradiology and artificial intelligence to allow patients to be treated remotely faster and at lower cost [224]. Audace Nakeshimana, founder and executive chairman of Insightiv, hopes that within 10 years, the organization can reach at least 10% of the undiagnosed population, which could help tens of thousands of patients get treated for cancer earlier. I spoke with Audace about how radiology report generation could help, and he said that they currently use ML for classification but their radiologists do not always find this helpful. They would like to focus on localization, perhaps starting with segmented classification but eventually moving toward "mini captions" for sub-regions in the image which the teleradiologists can interact with in their platform.

Companies like Insightiv are able to benefit from technical advancements because of organizations like Norrsken's Healthtech Hub Africa[8] program, which seeks to invest tens of millions of dollars into development and deployment of health tech. These kinds of initiatives allow impacted communities to build their own capacity for translating this kind of technology into solutions for their problems.

Embrace Pluralism. Relatedly, one common challenge throughout NLP — even in the general domain — is that the super-majority of non-translation work is done on English-only datasets. This can bias the solutions to reflect models which favor English sentence structures. One obvious opportunity for future work in non-English report generation is to use PadChest [45], whose text is in Spanish. Coincidentally to what Audace expressed interest in, PadChest's texts is not the full reports, like in MIMIC-CXR, but are instead "text snippet[s] extracted from the original report containing the radiographical interpretation." This offers yet another opportunity to distinguish itself from MIMIC-CXR and Open-I in report generation tasks.

Rethink Binaries and Hierarchies. Because structured classification tasks are easier to model than open-ended text generation (and because data is easier to deidentify and share), early ML for radiology work focused there [166]. However, the benefit of natural language over structured representation is that it allows for rich nuance (e.g., *"Projecting over the right costophrenic sinus, a soft tissue density has newly appeared. The density corresponds to an area of pleural thickening"*) and uncertainty (e.g., *"In addition, there is blunting of the left costophrenic sinus, so that the pleural effusion cannot be excluded."*) in the read. This less-structured task embraces the Data Feminist principle to rethink binaries and hierarchies; if some

[8] https://www.norrsken.org/hthubafrica

subset of disorders were categorized for classification, it would de facto push out solutions for patients that are harder to slot into a category.

Consider Context. One area where the task definition could improve is through the DF principle to consider context. Our model is formulated to generate a report from the input of one image. Depending on the use case intended, this may or may not be a useful research problem. It could be a simplified way to test the basic extraction of findings and description modules that could be part of a larger-scale effort for report generation. On the other hand, such a task would prove an inadequate substitute for the full radiologist workflow; the salience of certain findings in the radiology report may be influenced by the reason for exam, clinical history, and demographics. Additionally, the radiology study often has multiple images (frontal and lateral) as well as access to previous studies to compare against. Although some of the general concerns may be mitigated by working on the findings section instead of the impressions section, prior work found that nearly half of the reports examined in a few-hundred report selection of MIMIC-CXR included a comparison to the previous radiological study [34]. Without doing something to provide some of that context of the workflow, then the model would not be able to replicate the situated knowledge of the reports actually in the EHR. This is certainly one area of improvement for future work.

2.6.2 Project Impact Evaluation

As discussed in Section 2.6.1, this work is an early stage model on public, observational data to explore the feasibility of a given task. The “impact” of this project does not come from downstream metrics like lives improved or costs saved; it is research whose aim is to influence other research. As of February 2022, the paper describing this work was cited 89 times since being published in April 2019. To assess the impact of this work, I read through these works to understand how it has influenced future areas of research.

Of the 89 articles:

- 44 generate Chest X-Ray reports from images;

- 2 build upon techniques in the paper (evaluation metrics [34]; differentiable chexpert optimization [179]) without a dedicated focus on doing report generation, themselves;
- 10 were surveys, including of medical report generation [227, 182], medical ML broadly [251], and general domain NLP [129];
- 27 cite us in their related work sections, pointing to the work as representative of the "generating radiology reports" field;
- 5 were duplicate entries;
- and 1 was neither in English nor accessible inside the J-STAGE paywall.

Of the 44 works which generate reports, 16 compare their model against ours (1 got our code working [121], 1 re-implemented our model [188], 1 re-implements a subset of our model which they attribute to us [203], 13 reported the numbers straight from our paper). All 44 evaluate their models using standard natural language generation metrics (e.g., BLEU, CIDEr, METEOR) and 23 additionally use some other metric, such as CheXpert [121, 203, 192], MeSH [123], MIRQI [320], and manual clinician evaluation [167]. The most frequent non-NLG metric is CheXpert-based accuracy/precision/recall, which numerous works attribute to comparing against our work [151, 15, 192].

Evaluation. Many of the most popular natual language metrics like BLEU and CIDEr have been widely criticised [239]. Boag; et al. [35] demonstrated that these natural language generation metrics disagreed with a CheXpert-derived score about which generative model was better: a simple tri-gram decoder conditioned on a clinical context (i.e., correct but ungrammatical) or a randomly-selected report from the training set (i.e., grammatical but irrelevant to the image). The natural language metrics were rating the random model as better, preferring surface-level text similarity over clinical correctness. Given that level of discrepancy, how much faith should we be putting in these papers' evaluations, especially when many of the models score within a few points of each other? As will be discussed further in Chapter 3, Boag; et al. [34] cites this work in their pilot study investigating effective ways to evaluate Chest X-Ray report generation.

Optimizing for Clinical Coherence. One of the primary contributions of this work is incorporating domain knowledge into the objective function by using REINFORCE to optimize for the CheXpert-based Clinically Coherent Reward instead of just CIDEr. This contribution has been adopted by many other papers as well, using RL to improve the clinical correctness of generated report [205, 121, 192, 202, 204]. Xu et al. [313] extends the RL approach, adding a repetition penalty to encourage diversity in generated sentences to address the duplicate sentence issue we encounter. However, Nguyen et al. [202] observes, "reinforcement learning methods are often computationally expensive and practically difficult to convergence [sic]." Later works built upon this by predicting a continuous and differentiable version of CheXpert to allow for clinically coherent optimization without the technical complexities of RL [188, 169, 192, 179].

System Generalizability. In another improvement to our system, the very well-done work by Miura et al. [188] observes "Our work is most related to Liu et al. (2019); their system, however, is dependent on a rule-based information extraction system specifically created for Chest X-Ray reports and has limited robustness and generalizability to different domains within radiology." To address this, they did not use CheXpert, but instead implemented an RL model based on textual entailment (i.e., "contradiction detection") using Stanza [230], an open-source Python natural language processing toolkit supporting 66 human language. This approach would generalize more naturally both to non-radiology generation tasks and also non-English generation tasks (e.g, PadChest).

Bias We fail to measure performance with respect to different patient identities. In the year following when this work was published, Seyyed-Kalantari et al. [257] find state-of-the-art deep learning *classifiers* for X-Ray images are biased with respect to protected attributes. Although model performance bias is just one part of how machine learning can cause disparate harm, and bias analysis shouldn't be the only area of research focus for AI Ethics communities, it is nonetheless important to measure and address. In the 2021 FDA framework for regulating medical machine learning, the agency emphasizes "Because AI/ML systems are developed and trained using data from historical datasets, they are vulnerable to bias – and prone to mirroring biases present in the data. Health care delivery is known to vary by factors such as race, ethnicity, and socio-economic status; therefore, it is possible that biases present in our health care system may be inadvertently introduced into the algorithms" and that there is great need for "the identification and elimination of bias" [82]. This offers another potential opportunity of work which is as of yet addressed: are their biases in radiology reports, and would report generation models exacerbate any disparities?

Reproducibility Reproducibility is essential not only for works which would like to be the gold standard to compare against but also for all projects, especially ones working on public datasets. Indeed, the MIMIC-CXR data use agreement[9] stipulates that "any publication which makes use of the data will also make the relevant code available." This paper incorporated many technical sophistications: hierarchical LSTMs, attention, and reinforcement learning for discrete signals. Given the complexity of this system, such work would be very hard to re-implement from scratch. Although we did make the code available,[10] it is difficult to run. The github repo has no forks and only one star as of February 2022 (nearly three years post-publication). On top of typical dependency installation challenges, it has many configuration settings (e.g. unspecified environment variables) which need to be set as well as assumptions about data preprocessing and formatting, none of which are addressed in the (empty) README.

There is only one published work that was able to get this code to run [121]. Another published work re-implemented[11] our model themselves [188]. However, the 5 papers which compare their proposed models against ours [14, 280, 270, 320, 3] do so by reporting the paper's numbers directly (despite being from an alpha of MIMIC-CXR with different preprocessing, train/test split, and evaluation scripts), resulting in a comparison which is not apples-to-apples. Further, the implementation difficulties likely discourage additional direct comparisons, because over 15 works cite this paper and propose their own models but do not compare against our model,

[9] https://physionet.org/content/mimic-cxr/2.0.0

[10] https://github.com/stmharry/interpretable-report-gen

[11] https://github.com/ysmiura/ifcc/blob/master/clinicgen/models/cnnrnnrnn.py

choosing instead to compare against other models whose source code is available, such as R2Gen [57], TopDown [8], TieNet [300], CoAtt [131], etc.

Chapter 3

Evaluating Quality of Generated Text

In Chapter 2, I described a successful effort to generate state-of-the-art radiology reports automatically. This was done by hierarchically generating topics from images, then sentences from topics. The final system is also optimized with reinforcement learning for both readability (via CIDEr) and clinical correctness (via the novel Clinically Coherent Reward). Our system outperformed a variety of compelling baseline methods across readability and clinical efficacy metrics on both MIMIC-CXR and Open-I datasets.

However, as was briefly discussed earlier, evaluating the quality of generated text is very difficult. Researchers have struggled to validate their NLG evaluations in both the general domain [207, 145] and clinical domain [35]. The gold standard of a report's "good-ness" would be how well it improves outcomes for the patient or hospital: perhaps it catches more illnesses than a bad report would, or perhaps it saves time/money for hospital operations. Of course, one cannot run a randomized controlled trial because a bad model would result in significant harm to patients. That is why the clinical domain needs an appropriate metric to serve as a proxy when the true outcome itself cannot be obtained. The difficult challenge is in determining whether a proxy is appropriate; BLEU is a proxy, but the field has spent over a decade pointing out the many flaws it has.

In this chapter, I describe a long-term project to characterize what makes clinical text uniquely challenging and make numerous attempts to evaluate the quality of automatically-generated radiology reports.

This work was done in collaborations with:

- Section 3.1: Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alsentzer, and Peter Szolovits [168].
- Section 3.2: Hassan Kané, Saumya Rawat, Jesse Wei, and Alexander Goehler [34].
- Section 3.3: Hassan Kané, Saumya Rawat, Alexander Goehler, Vikram Venkatesh, and Patricia Balcacer.

3.1 Flaws in Standard Evaluation Metrics

Existing metrics in the general domain can be broadly categorized into using n-gram matching, embedding matching, or learned functions. There are also domain-specific metrics for clinical tasks using basic information extraction tools.

The most commonly used metrics for text generation count the number of n-grams that occur in the reference and candidate text. BLEU and ROUGE are the most widely used metric in machine translation [220, 164]. They leverage precision and recall over different values of n-gram overlaps to obtain a final score. METEOR relaxes the n-gram overlap approaches and enables synonyms overlap [19]. In addition to BLEU, ROUGE and METEOR, we also included more recently introduced metrics based on Transformer architectures. These include BERTScore, NUBIA and BLEURT [319, 255, 140].

Specific to the domain of radiology report generation, CheXpert uses rules-based approaches to identify the presence or absence of 14 specific diagnoses [127]. Metrics can be derived by extracting the diagnostic information from the reference and candidate reports and computing agreement (e.g., precision, recall, accuracy, etc.).

Evaluation metrics are meant to be proxies for bepoke human evaluations, therefore any new metric that is developed is measured to demonstrate its correlation with some form of human judgment. The original BLEU paper had human evaluators score candidate translations from 1 (very bad) to 5 (very good) and then demonstrated that BLEU ranks 5 systems in the same order that human annotators do [220]. CIDEr was created in 2014 for image captioning, and became very popular for benchamrking on the MS COCO dataset [289]. CIDEr annotators were asked to decide which of two candidate captions better agrees with a reference caption, and then CIDEr was shown to agree with the annotator rankings more strongly than previous metrics like BLEU and METEOR. Several papers have proposed moving away from correlation and more towards multidimensional evaluation where sentence corruptions are introduced as "unit tests" for how well the metric responds [32, 141].

Many of the most popular metrics have been widely criticised [239]. In the context of text simplification, BLEU has a very weak – and in some cases, negative – correlation with human judgment on grammaticality, meaning preservation, and simplicity [277]. Even in the context of machine translation (for which it was originally created), BLEU correlates poorly with human judgment on both adequacy (i.e., whether the hypothesis sentence fully captures the meaning of the reference sentence) and on fluency (i.e., the quality of language in the hypothesis sentence) [49].

These concerns are likely especially true in the clinical domain, where we care not only about free-text readability, but also about the accuracy of the stated clinical conclusions. Further, these metrics were designed to be all-purpose tools, independent of any domain, which limits how reliable they might be expected to be for a highly specialized area such as medicine. It may prove true that these tools are sufficient proxies for even doctor judgment, but that has not yet been shown – these metrics were validated based on correlation with human judgment on generic sentences with a large number of reference sentences.

3.1.1 Inconsistent Ranking of Methods

In Chapter 2, we evaluated the generative models using both NLG and CheXpert-based metrics. In this subsection, I demonstrate that such a combination is insufficient for determining the quality of generated text. I do this through a not-quite-toy, but simplified, scenario where the metrics are used to evaluate some baselines. I find that the NLG and CheXpert-based metrics disagree on which models are better, even when the outcome should be intuitively clear from knowledge about the baseline

Table 3.1: Automatic evaluation metrics of baseline methods for image captioning task.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	CheXpert Accuracy	CheXpert Precision	CheXpert F1
Random	**0.265**	**0.137**	**0.070**	**0.036**	**0.570**	0.770	0.146	0.148
3-gram	0.206	0.107	0.057	0.031	0.435	**0.782**	**0.225**	**0.185**

approaches.

This work uses the MIMIC-CXR dataset, which consists of chest x-ray images and reports from 63,478 patients. We subdivide the data into a train set of 75,147 and a test set of 19,825 images, with no overlap of patients between the two. Radiological reports are parsed into sections and we use the findings section. The baseline models are as follows:

- **Random Retrieval** (Random): Ignore the query image, and instead draw a random report from the training set as the "generated" text.
- **3-gram Language Model** (3-gram): Identify the 100 train images that are closest to the query image (in a CNN-extracted 1024-dim space). Learn a 3-gram language model from their reports. Sample from that model to generate a report for the query image.

We would expect the random retrieval baseline's reports would be readable, but not relevant to the query image at all; as such, they would be unlikely to score well either on the text-generation metrics or on our measures of clinical relevance.

Table 3.1 shows the results of this experiment across the NLG and CheXpert metrics. As expected, the random sentences score lower than the 3-gram model on the clinical correctness metric, because the 3-gram model samples from similar cases. On the other hand, the random sentences model unexpectedly scores higher than the 3-gram model on standard NLG tasks. In retrospect, this makes sense because the NLG metrics are essentially looking at simple n-gram overlap with template-heavy reference reports. Although the random sentences may be completely irrelevant to the query image, it is undoubtedly true they look just like how a report is supposed to look.

This should give some pause about what "good" performance on these tasks looks like. Standard NLG metrics are ill-equipped to measure the quality of clinical text. The over-reliance on n-gram overlap causes these metrics to favor irrelevant-but-fluent reports over correct-but-ungrammatical ones. The code for this work is publicly available.[1]

3.1.2 Challenges of the Clinical Domain

Although these metrics have been used in evaluating radiology report generation [168], they are often designed for other contexts with underlying assumptions about the number of reference sentences available as well as the complexity of the sentences to be analyzed. When CIDEr was introduced, its authors demonstrated that "humans and CIDEr agree with a high correlation," but they did so when there were 20-50 reference captions per images as well as very low-complexity image captions (e.g., "a cow is standing in a field"). It is not clear whether these findings would hold in the clinical domain, where there is: only one reference report, many more tokens, and a strong emphasis on factual correctness.

To better understand the differences between the simple general domain image descriptions and clinical text, we use standard readability scores to assess the complexity of a given piece of text. The Dale–Chall readability formula and Gunning-Fog index measures the years of formal education a person needs to understand the text on the first reading [53, 107]. In the case of Gunning-Fog, the score is meant to directly indicate the number of school years (e.g., 7 means 7th grade, 12 means 12th grade, etc) and Dale-Chall works similarly but on a 1-10 scale. The Flesh readability score rates documents on a 100-point scale based on the number of words and sentence and syllables per word [85]. Unlike the previous two indices, higher Flesch scores indicate easier-to-read documents.

Table 3.2 demonstrates many of the differences between PASCAL-50S (a dataset introduced in the CIDEr paper) and MIMIC-CXR. We observe that PASCAL-50S indeed has 50 reference reports, each of which has 5 times fewer tokens than a MIMIC-

[1] `https://github.com/wboag/cxr-baselines`

Table 3.2: Linguistic characteristics of PASCAL-50S (CIDEr's annotations) and the MIMIC-CXR radiology reports.

Average Characteristic	PASCAL-50S	MIMIC-CXR
number of references (per image)	50.00 ± 0.0	1.00 ± 0.0
sentence count (per reference)	1.00 ± 0.0	5.29 ± 1.9
word count (per reference)	9.82 ± 3.2	55.25 ± 25.2
Dale–Chall readability score (per reference)	5.23 ± 3.2	9.61 ± 1.0
Gunning-Fog index (per reference)	11.20 ± 4.7	20.06 ± 2.8
Flesch readability score (per reference)	96.08 ± 15.6	63.26 ± 12.5

CXR report. Additionally, the Dale-Chall and Gunning-Fox readability scores suggest that nearly twice as many yeears of formal education are required to understand radiology reports than simple image descriptions. Clinical text is demonstrably more complicated than the general domain text that previous metrics were developed for.

3.2 Attempt 1: Pilot Study in Clinician Judgment

Upon realizing metrics like BLEU and CIDEr were especially too unreliable for the clinical domain (where clinical correctness is critical), I brought together machine learning experts with radiologists to co-design a better evaluation metric. The interdisciplinary collaborative process involved outreach, multiple interviews beforehand, a pilot annotation process, and an exit reflection with the radiologist annotators.

The long-term goal is to eventually design a better evaluation metric for determining whether an automatically-generated radiology report is good. However, the challenges involved would likely not be solved on the first attempt. In this pilot study, we provide three main contributions:

- We develop and conduct an annotation task to collect clinical judgment on 400 candidate reports from 100 radiology images. This provides much-needed guidance on what makes a report good or bad.
- We examine what radiologists look at when evaluating a report. This is useful both to understand the current limitations of the task framing and also to help

inform future development of a better evaluation metric for Chest X-Ray report generation.

- We demonstrate some of the outreach tools we used for initial contact with domain experts to help create discussions and eventual partnerships.

Our findings highlight the need for data scientists to work closely with clinical experts to build meaningful tasks and models.

3.2.1 Methodology

This collaborative effort was done in two parts: qualitative discussions to design the task, followed by analysis. We conducted interviews with radiologists from three hospitals (two from Boston, MA and one from Atlanta, GA). After these conversations, we collected annotations from 2 radiologists and analyzed the results.

Designing the Annotation Task

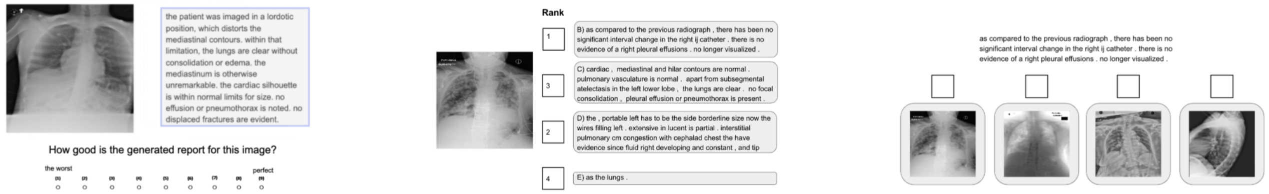
Based on prior work in radiology report generation [35] and evaluation metric creation [220, 289], we had a rough idea of the collaboration and data collection approach that we had in mind: radiologists need to read generated reports and decide (in some way, shape, or form) whether they are good or not.

The simplest way to go about this could be to display an image + report and ask the radiologist to rank how good it is from 1-to-5. Unfortunately, this approach suffers from broader design issues: behavioral economics demonstrates that humans can be inconsistent. We can see an example of this from the Sentences Involving Composition Knowledge (SICK) dataset [177], where prior work observes that the same kind of sentence transformations can be scored inconsistently [32]:

1. A man is holding a frog had a 2.1/5 similarity with
   There is no man holding a frog.
2. A man is playing soccer had a 4.8/5 similarity with
   There is no man playing soccer.

Although some of these annotator calibration concerns can be solved with mean nor-

Figure 3-1: Three different annotation tasks we considered for the radiologists to perform.



the patient was imaged in a lordotic position, which distorts the mediastinal contours. within that limitation, the lungs are clear without consolidation or edema. the mediastinum is otherwise unremarkable. the cardiac silhouette is within normal limits for size. no effusion or pneumothorax is noted. no displaced fractures are evident.

How good is the generated report for this image?

the worst (1) ○ (2) ○ (3) ○ (4) ○ (5) ○ (6) ○ (7) ○ (8) ○ (9) ○ perfect

Rank

1 B) as compared to the previous radiograph , there has been no significant interval change in the right ij catheter . there is no evidence of a right pleural effusions . no longer visualized .

3 C) cardiac , mediastinal and hilar contours are normal . pulmonary vasculature is normal . apart from subsegmental atelectasis in the left lower lobe , the lungs are clear . no focal consolidation , pleural effusion or pneumothorax is present .

2 D) the , portable left has to be the side borderline size now the wires filling left . extensive in lucent is partial . interstitial pulmonary cm congestion with cephalad chest the have evidence since fluid right developing and constant , and tip

4 E) as the lungs .

as compared to the previous radiograph , there has been no significant interval change in the right ij catheter . there is no evidence of a right pleural effusions . no longer visualized .

□ □ □ □

(A) **Direct Assessment**. Radiologist would be asked to select how good the generated caption is for the image from 1-10.

(B) **Caption Ranking**. Radiologist would be asked to rank 4 proposed captions based on how well each describes the given image.

(C) **Image Selection**. Radiologist would be asked to select which image is the one being described by a given caption.

malization, ensuring that a particularly harsh annotator doesn't distort the average, the larger problem is that annotators are not only inconsistent with one another, but they can also be inconsistent with themselves depending on their context and priming [283].

With this in mind, we explored a few potential ways to pose the annotation task for doctors. Figure 3-1 demonstrates a few ways to ask annotators to make judgments, such as a ranking-based approach (3-1b) or image selection (3-1c).

Interviews with Radiologists

In order to reach out to radiologists to discuss this project, we created a "1-pager" to send to them before our call, inspired by the *Collabsheets* list of "simple" questions for computer scientists and clinicians to discuss [250]. The 1-pager is shown in Figure 3-2, and its purpose is to give a background on where we are coming from and focus the conversation on the kinds of questions that seemed important to us.

On many questions, there was a strong consensus among the radiologists. They all agreed that clinical correctness is the most important factor in determining whether a report is good. Additionally, each radiologist talked about their field's move towards more structured, templated reports, with some suggesting that perhaps an evaluation metric should try to favor regularity. Finally, there was overall agreement that for an annotation task like this (where they were not being asked to write their own reports) it might be nice to have a DICOM image viewer that could allow them to zoom, adjust contrast, etc., but such functionality would not be necessary.

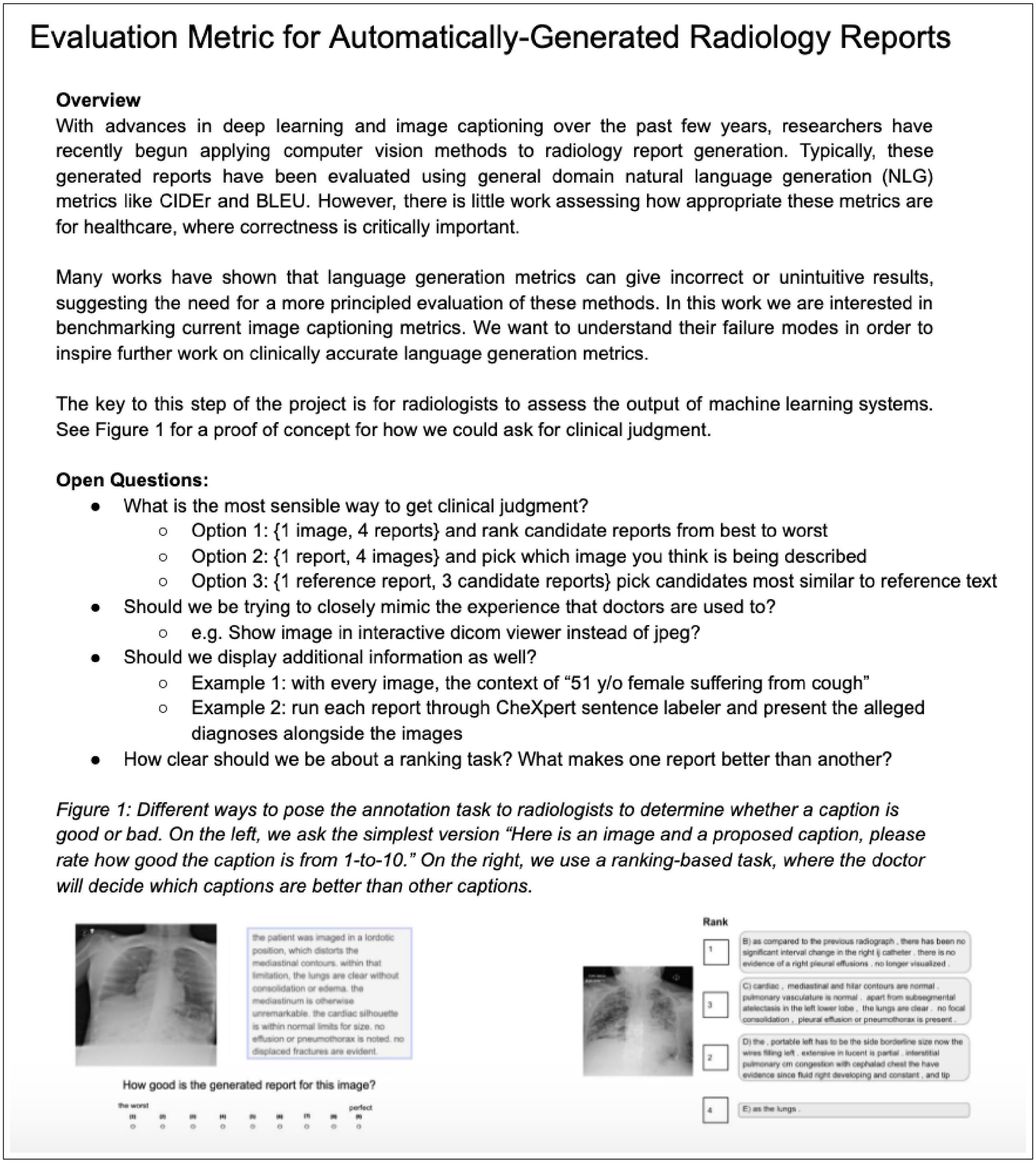
During the course of the interviews, there were a few other concepts raised by radiologists which we had not considered when designing the 1-pager in Figure 3-2, including:

- Many images are simply “normal heart, normal lungs, etc.” We should purposefully select a diversity of diagnoses in the annotation set.
- When designing a metric eventually, it may be useful to look for words conveying levels of uncertainty (e.g., “consistent with” vs. “suggests”).

There was, however, some disagreement among the clinicians. Interestingly, the notion that different doctors could disagree on healthcare expert opinions was surprising for some computer scientists on the team. As an analogy: doctors can disagree on report structure and evaluation in the same way that computer scientists disagree over the promise vs. hype of different deep learning methods. No field is a monolith.

One radiologist questioned whether any of the proposed annotation tasks (direct assessment, caption ranking, and image selection) were the most meaningful thing to measure. They suggested perhaps we should create an interface where the generated report is a “first draft” for the annotator to modify until they are satisfied with the final product. This would, however, be a much more involved undertaking for our annotators. Additionally, radiologists disagreed over whether it was worth including background information about the patient (e.g., “51 y/o female suffering from cough”).

Figure 3-2: This “1-pager” document was sent to radiologists during outreach when setting up initial conversations about this project’s goals.



Evaluation Metric for Automatically-Generated Radiology Reports

Overview

With advances in deep learning and image captioning over the past few years, researchers have recently begun applying computer vision methods to radiology report generation. Typically, these generated reports have been evaluated using general domain natural language generation (NLG) metrics like CIDEr and BLEU. However, there is little work assessing how appropriate these metrics are for healthcare, where correctness is critically important.

Many works have shown that language generation metrics can give incorrect or unintuitive results, suggesting the need for a more principled evaluation of these methods. In this work we are interested in benchmarking current image captioning metrics. We want to understand their failure modes in order to inspire further work on clinically accurate language generation metrics.

The key to this step of the project is for radiologists to assess the output of machine learning systems. See Figure 1 for a proof of concept for how we could ask for clinical judgment.

Open Questions:

- What is the most sensible way to get clinical judgment?
  - Option 1: {1 image, 4 reports} and rank candidate reports from best to worst
  - Option 2: {1 report, 4 images} and pick which image you think is being described
  - Option 3: {1 reference report, 3 candidate reports} pick candidates most similar to reference text
- Should we be trying to closely mimic the experience that doctors are used to?
  - e.g. Show image in interactive dicom viewer instead of jpeg?
- Should we display additional information as well?
  - Example 1: with every image, the context of “51 y/o female suffering from cough”
  - Example 2: run each report through CheXpert sentence labeler and present the alleged diagnoses alongside the images
- How clear should we be about a ranking task? What makes one report better than another?

Figure 1: Different ways to pose the annotation task to radiologists to determine whether a caption is good or bad. On the left, we ask the simplest version “Here is an image and a proposed caption, please rate how good the caption is from 1-to-10.” On the right, we use a ranking-based task, where the doctor will decide which captions are better than other captions.

the patient was imaged in a lordotic position, which distorts the mediastinal contours. within that limitation, the lungs are clear without consolidation or edema. the mediastinum is otherwise unremarkable. the cardiac silhouette is within normal limits for size. no effusion or pneumothorax is noted. no displaced fractures are evident.

How good is the generated report for this image?

the worst (1) ○ (2) ○ (3) ○ (4) ○ (5) ○ (6) ○ (7) ○ (8) ○ (9) ○ perfect

Rank

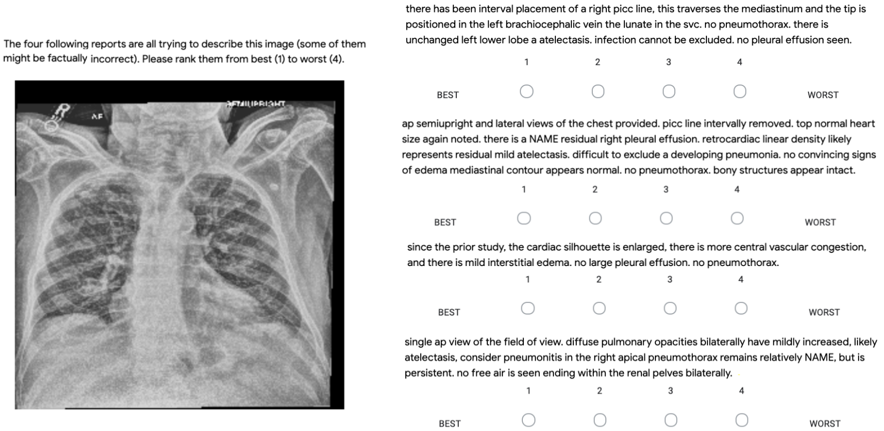
1 B) as compared to the previous radiograph , there has been no significant interval change in the right ij catheter . there is no evidence of a right pleural effusions , no longer visualized .

3 C) cardiac , mediastinal and hilar contours are normal . pulmonary vasculature is normal . apart from subsegmental atelectasis in the left lower lobe , the lungs are clear . no focal consolidation , pleural effusion or pneumothorax is present .

2 D) the , portable left has to be the side borderline size now the wires filling left . extensive in lucent is partial . interstitial pulmonary cm congestion with cephalad chest the have evidence since fluid right developing and constant , and tip

4 E) as the lungs .

Figure 3-3: An example instance of the chosen annotation task.



The four following reports are all trying to describe this image (some of them might be factually incorrect). Please rank them from best (1) to worst (4).

there has been interval placement of a right picc line, this traverses the mediastinum and the tip is positioned in the left brachiocephalic vein the lunate in the svc. no pneumothorax. there is unchanged left lower lobe a atelectasis. infection cannot be excluded. no pleural effusion seen.

1 2 3 4

BEST ○ ○ ○ ○ WORST

ap semiupright and lateral views of the chest provided. picc line intervally removed. top normal heart size again noted. there is a NAME residual right pleural effusion. retrocardiac linear density likely represents residual mild atelectasis. difficult to exclude a developing pneumonia. no convincing signs of edema mediastinal contour appears normal. no pneumothorax. bony structures appear intact.

1 2 3 4

BEST ○ ○ ○ ○ WORST

since the prior study, the cardiac silhouette is enlarged, there is more central vascular congestion, and there is mild interstitial edema. no large pleural effusion. no pneumothorax.

1 2 3 4

BEST ○ ○ ○ ○ WORST

single ap view of the field of view. diffuse pulmonary opacities bilaterally have mildly increased, likely atelectasis, consider pneumonitis in the right apical pneumothorax remains relatively NAME, but is persistent. no free air is seen ending within the renal pelves bilaterally.

1 2 3 4

BEST ○ ○ ○ ○ WORST

Pilot Study: Annotation Task

Based on the feedback from initial conversations, we conducted a data collection pilot study. Two radiologists annotated 100 images (400 captions) apiece. Figure 3-3 demonstrates one instance of this task: for a given image, radiologists needed to rank 4 possible reports based on how well they describe the findings of the image. Each radiologist viewed the same images in the same order.

For each image, we presented the annotator with the following statements:

1. “The four following reports are all trying to describe this image (some of them might be factually incorrect). Please rank them from best (1) to worst (4).”
2. “Briefly describe how you arrived at this ordering (a few simple bullet points is fine)”
3. “Confidence that another radiologist would arrive at the same choice for best report (1=Not confident at all, 5=Very confident)”

For each image, we generate four different reports using the following methods: reference, 3-gram, nearest neighbor (1-NN), and random-report. The “reference” report refers to the actual report written by the radiologist, logged in the EHR. The nearest neighbor (1-NN) report is produced by returning the report of the closest image (in the DenseNet-induced feature space) training set. Similarly, the “random-report” is the associated report of a *randomly* selected image from the training set. Finally, the “3-gram” is produced by retrieving the 100 closest images and fitting a tri-gram model from their associated reports.

Performance of Evaluation Metrics

Ranking the four reports for a single image produces 6 comparisons (i.e., best > the other three, the second best > the other two, the third best > the worst), though 3 of those comparisons involve the reference sentence. To determine how strongly a given metric agrees with radiologist judgment, we compute the specific metric score for each of the 3 non-reference candidates[2] and determine the number of pairwise comparisons where the metric agrees with the experts. When the two experts did not agree on a pairwise comparison, that ranking was excluded.

For the metrics, we evaluate many evaluation metrics, including baselines (random-score, length), readability scores (Dale-Chall), n-gram (BLEU, CIDEr), embedding (BERTScore), and CheXpert accuracy.

3.2.2 Results

For 100 images, ranking 4 reports results in 600 binary comparisons. Of those 600 comparisons, the annotators agreed with each other on 459 (i.e., 76.5% of the time).[3] Of the 300 rankings which did not include the reference report, radiologists agreed on 199 rankings (i.e., 66% of the time).

In line with prior work [35], the clinical correctness of the 3-gram model (0.353) is higher than the random-report model (0.319) but the nearest neighbor achieves the highest level (0.437).

Table 3.3 shows how often each metric agreed with consensus radiologist rankings. We include the "random-score" metric (not to be confused with the "random-report" method) as a sanity check: if the metric were randomly assigning numbers, it would get the ranking correct 50% of the time. The "Percent Ties" column denotes how often

[2] We do not compute the metric on the reference because, by definition, it would score 100% as you would be comparing the reference against itself.

[3] There were 5 data entry errors where two reports were given the same ranking (e.g., ranking of 1,2,3,3 or 1,2,4,4) even though the task did not allow for ties. For those five entries, the authors used the given explanations to infer what the annotator meant and broke the ties.

Table 3.3: Of the 199 consensus comparisons, how often would each metric rank the two reports the same way the radiologists did?

Metric	Percent Agree	Percent Ties
random-score	50.0%	0
choose shorter report	54.3%	0.5%
Dale-Chall Readability Index	58.3%	0
BLEU-1	53.3%	0
BLEU-4	50.8%	0
ROUGE-1	56.3%	1%
CIDEr	58.8%	0
BERTScore	**61.3%**	0
chexpert-accuracy	43.7%	24.6%
chexpert-accuracy + .001*CIDEr	57.3%	0.5%

Table 3.4: Top n-grams from the explanations provided by annotators for decision-making. Phrases containing stop words were removed.

unigram	Count
“factually”	16
“all”	17
“wrong”	18
“not”	21
“correct”	24

bigram	Count
“even though”	6
“most correct”	7
“hard to”	9
“factually wrong”	10
“all but”	13

trigram	Count
“are factually wrong”	3
“one and two”	3
“not sure if”	4
“all but one”	5
“is hard to”	6

a given metric was not able to pick either report (e.g., if two reports were each correct on 9/14 findings, then chexpert-accuracy would be tied at 64% a piece). Because this is so common for chexpert-accuracy metrics, we also report how well it would perform when CIDEr is used to break the ties (i.e. *+ .001*CIDEr*) BERTScore attains the top performance of 61.3%; CheXpert only achieves 57.3%.

Self-Reported Annotator Rationale

We were curious to understand how radiologists would decide to rank reports. Would they focus on style, factual correctness, grammar, concision, or potentially other factors?

One of the annotators qualitatively described their process for completing the ranking task in an interview. They made it clear that the top criterion is factual correctness: "It doesn't matter how nice or brief a report is. If it's factually wrong, then it's bad." To rank the candidates, they would do an initial pass to group reports into two buckets: "plausible" and "wrong." From there, they would look at each bucket and identify which errors were more egregious (e.g., the report with a rare type of error was ranked worse than a report with a common type of error). Whenever two reports were both plausible without any disqualifyingly bad mistakes, they would look to see which one was more complete, especially since some omissions (e.g., failure to mention a lung lesion) would be more glaring than others.

Based on the quantitative results, the other annotator seemed to agree about the importance of correctness. After each image's ranking, annotators were asked to "Briefly Describe How You Arrived at This Ordering (a few simple bullet points is fine)." Table 3.4 depicts the top-5 most frequent unigrams, bigrams, and trigrams of the rationales experts gave in response. We can see through uses of phrases like "correct", "wrong", "factually", and "are factually wrong" that they are explaining their decisions as decisions of factual correctness. Additionally, they convey the challenges of comparing two non-perfect candidates through phrases like "most correct" and "all but one."

In both the qualitative and quantitative analyses, there was little to no discussion of readability or grammaticality.

Why Do Radiologists and CheXpert Disagree on 3-grams?

One surprising part about these "factual correctness"-based explanations is that Table 3.3 shows the Dale-Chall Readability Index agreed with the radiologists (58.3%) more often than any of the CheXpert-based metrics (57.3%). This finding continues the discrepancy initially discovered in [35] where 3-gram outperformed the random-report model on CheXpert's clinical correctness but underperformed on standard evaluation metrics [35].

We can see in Table 3.5 that ngram-based reports tend to have higher variance of readability, suggesting there may be especially difficult-to-read reports which are

Table 3.5: Readability scores of the 100 ngram-generated reports vs. the 100 random-report candidates.

Average Characteristic	random-report	3-gram
Dale–Chall readability score	9.61 ± 0.9	9.10 ± 1.5
Gunning-Fog index	20.03 ± 2.6	19.2 ± 3.7
Flesch readability score	63.66 ± 13.0	65.90 ± 17.3

still coherent enough for CheXpert's rules to parse correctly. It may be the case that especially ungrammatical reports came across as "non-sensical" to annotaors, which *could* be considered part of "correctness."

Additionally, many reports reference previous x-ray images, some directly (e.g., "prior study" or "picc line removed") and others subtly (e.g., "interval placement of a right picc line" and "have mildly increased"). This suggests that the single image alone might not contain enough information to assess whether a report is correct or not. Of the 400 reports shown to annotators, 231 of them contain a mention of at least one of the following:

- "previous"
- "comparison"
- "compared"
- "prior"
- "from DATE"
- "unchanged"

Relatedly, 54/400 allege the existence of a lateral radiograph to accompany the AP view. These "missing inputs" make it difficult for annotators to correctly assess the quality of radiology reports, and suggest the report generation will require more than fitting generative models to map from an EHR's stored image to its accompanying report.

3.3 Attempt 2: Multi-Dimensional Evaluation

The pilot study had shortcomings which prevented its collected data from being as useful as was initially hoped. However, we incorporated some lessons from that effort into a second annotation attempt to measure the quality of generated radiology reports.

Once again, we found that this second attempt additionally did not produce final data which was as valuable as initially hoped. The aim of this section is to once again help others learn from the design and analysis of this case study. In particular, our contributions are as follows:

- We build upon our work framing the annotation task in a way that would be useful.
- We collect annotations for 501 reports from 167 images.
- We demonstrate the challenges that arose in working with an interdisciplinary team.
- We present analysis which is able to interrogate and "sanity check" collected data to self-assess its consistency and value.

3.3.1 Methodology

First, we motivate an improved way to assess text quality by separating different concepts (e.g., correctness, grammar, etc) into different dimensions of measurement. We then reflect on shortcomings of the previous attempt and describe the collaborative process of working with radiologists to design a second attempt. Finally, we analyze the results of the radiologist annotations.

Designing the Annotation Task

One shortcoming of the first attempt was that it was sample inefficient; annotators would essentially read two full reports and give 1 bit of information ("which one is better?"). Although we had additionally requested the annotators give written explanations for their decisions, the rationales appeared to not tell the whole story (e.g., claimed factual correctness as most important factor, despite text simplicity scores correlating higher with judgment than CheXpert information extraction) and were difficult to test because of their unstructured nature. A second shortcoming of the ranking-based approach is that by wrapping all decision-making into one binary choice, it obscured the multiple kinds of ways a report could be deficient. This makes it difficult disaggregate false positives vs. false negatives vs. style vs. brevity, etc. Finally, the radiologists we spoke preferred to be able to rate things on a scale (e.g., 1-5, 1-7).

Existing automated metrics such as BLEU, CIDEr, and even CheXpert, have struggled to sufficiently serve as useful indicators of success in report generation. One main reason for this is that there are many different ways that a report can be wrong, and any one 0-1 scale will be too reductive. Consider the following three reference-candidate pairs:

1. "There is a man playing a guitar" vs.
   "There is no man playing a guitar"
2. "There is a man playing a guitar" vs.
   "There is a guitar playing a man"
3. "There is a man playing a guitar" vs.
   "There is a rabbit eating a flower"

Which of these pairs should have the lowest similarity? The sentences in the first pair — despite having the same subject, object, etc — are semantically opposite in meaning. In the second pair, the exact same set of words are used, but the second sentence is absurd and has a different meaning. Finally, in the third sentence, they aren't even in the same topic. This example illustrates the challenge in coming up with a scheme which satisfyingly crams all forms of similarity into one axis of quality.

To address this, our second attempt adopted a multi-dimensional assessment of: factual correctness, comprehensiveness, style, and overall quality. Emphasizing the multiple different aspects of quality both serves a diagnostic role (i.e., could help research teams debug models' shortcomings) and also facilitates conversations around values and requirements for the intended use case. For instance, if a model is the "last line of defense" then it must not miss any important findings, whereas if a model is looking for high priority cases that can "cut the line" but the default process is the standard of care, then you should work to avoid false positives and "alarm fatigue."

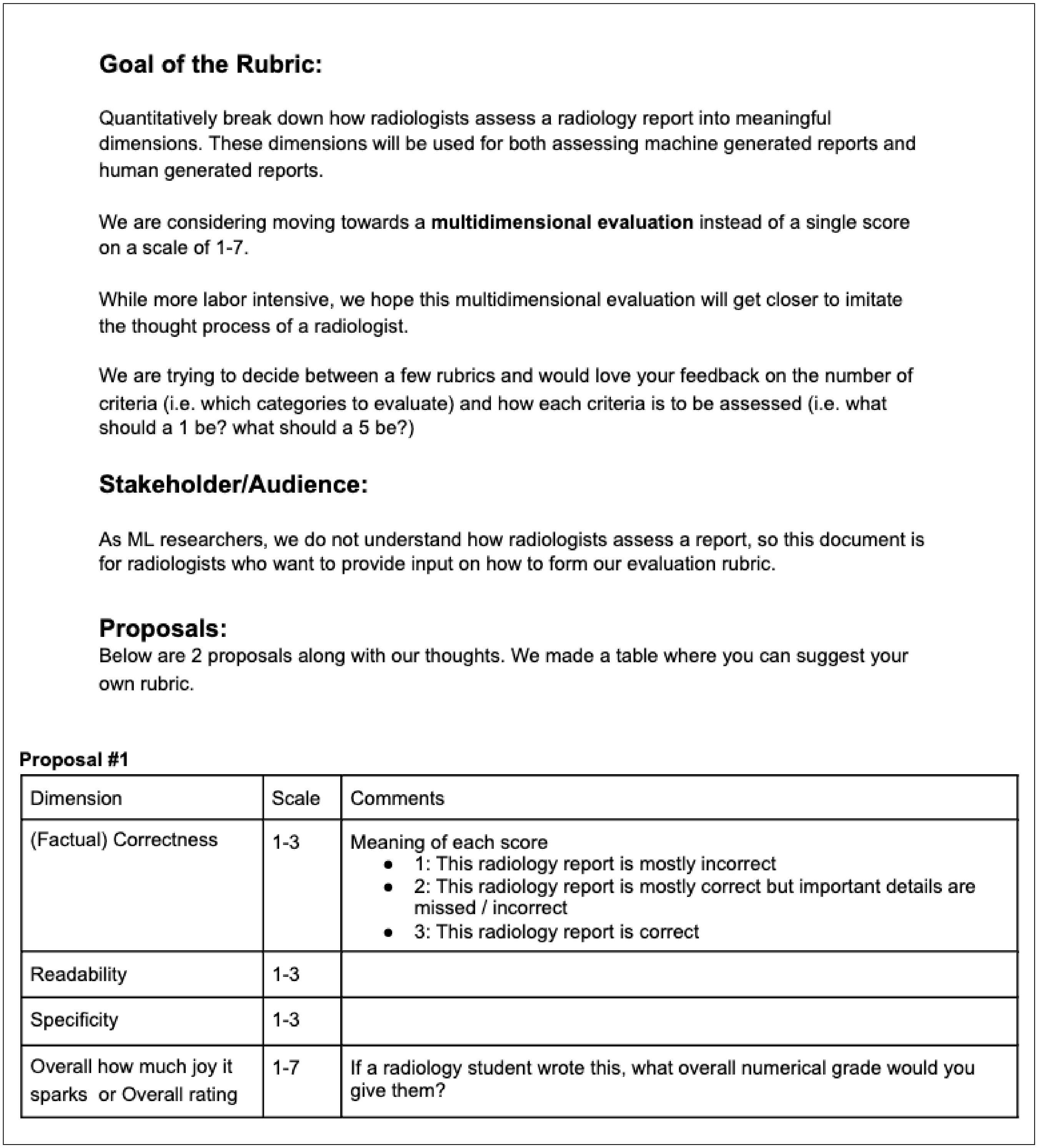
Interviews with Radiologists

To demonstrate the value in multiple dimensions of evaluation, we worked as an interdisciplinary team of computer scientists and radiologists to define meaningfully different criteria: factual correctness, comprehensiveness, and presentation/style. We created a worksheet to facilitate communication between the computer scientists and radiologists on the team, which can be seen in Figures 3-4 and 3-5.

Radiologists had numerous interesting suggestions, including:

- It was not intuitive for them to try to evaluate separably different dimensions of a report was not, because that is not what they were trained to do.
- They worried that the original report would be so much better than all of the generated reports that it could bias the annotator in choosing the 2nd best through 4th best report because by comparison they all look very bad.
- It would be helpful to have a PACS-like dicom viewer so that they could inspect the images much more closely than last time.
- Unlike in the pilot (which was a google form that required submitting all of the annotations at once), it would be better to use a google sheet, which saved partial submissions so that busy radiologist annotators could spread their work out across multiple sittings.

Figure 3-4: We iteratively filled in this 2-page document (1 of 2) between meetings with our radiologist collaborator. Walking through it together helped build a shared understanding with them about what was important to measure and why.



Goal of the Rubric:

Quantitatively break down how radiologists assess a radiology report into meaningful dimensions. These dimensions will be used for both assessing machine generated reports and human generated reports.

We are considering moving towards a **multidimensional evaluation** instead of a single score on a scale of 1-7.

While more labor intensive, we hope this multidimensional evaluation will get closer to imitate the thought process of a radiologist.

We are trying to decide between a few rubrics and would love your feedback on the number of criteria (i.e. which categories to evaluate) and how each criteria is to be assessed (i.e. what should a 1 be? what should a 5 be?)

Stakeholder/Audience:

As ML researchers, we do not understand how radiologists assess a report, so this document is for radiologists who want to provide input on how to form our evaluation rubric.

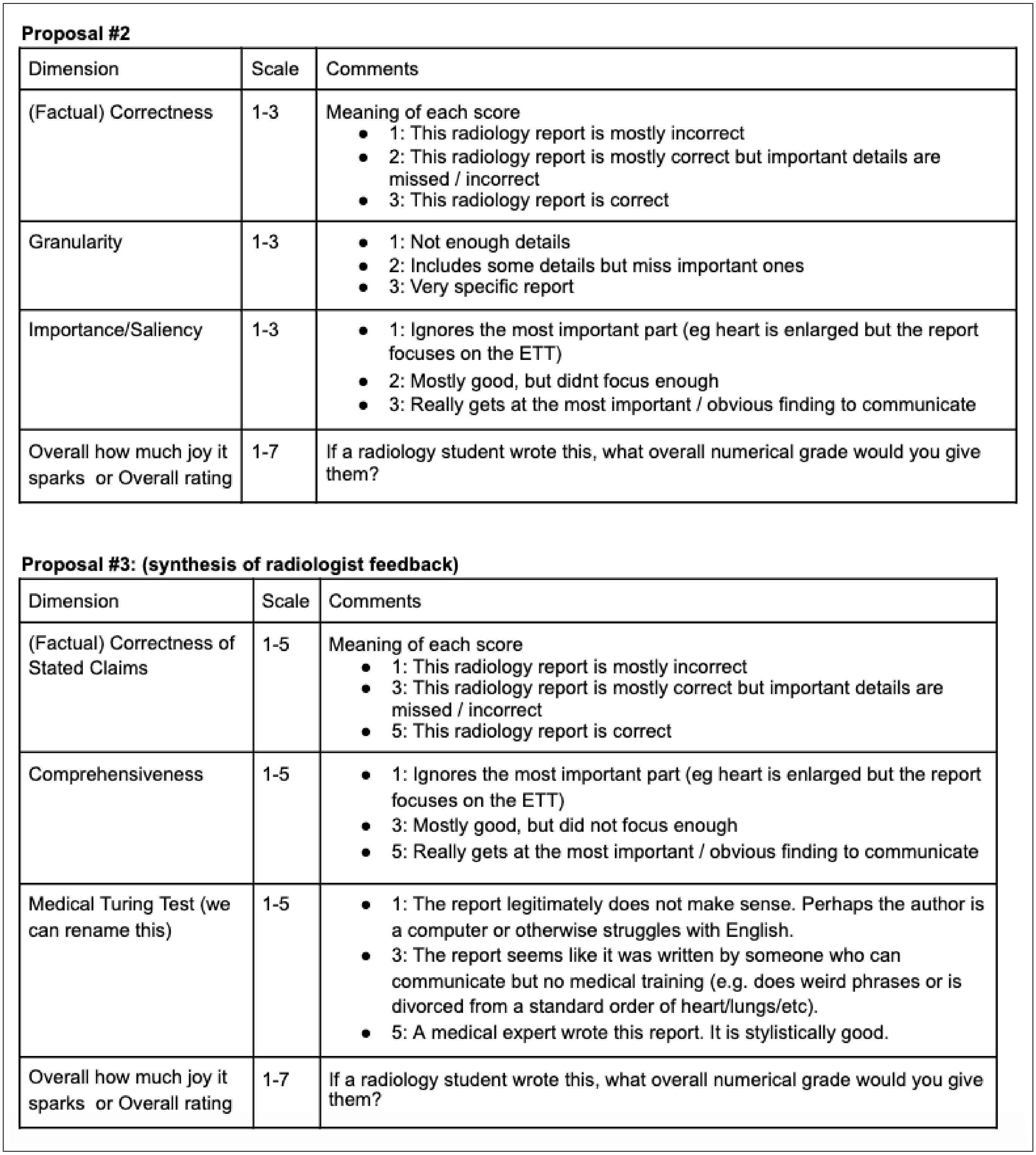
Proposals:

Below are 2 proposals along with our thoughts. We made a table where you can suggest your own rubric.

Proposal #1

Dimension	Scale	Comments
(Factual) Correctness	1-3	Meaning of each score • 1: This radiology report is mostly incorrect • 2: This radiology report is mostly correct but important details are missed / incorrect • 3: This radiology report is correct
Readability	1-3	
Specificity	1-3	
Overall how much joy it sparks or Overall rating	1-7	If a radiology student wrote this, what overall numerical grade would you give them?

Figure 3-5: We iteratively filled in this 2-page document (2 of 2) between meetings with our radiologist collaborator.



Proposal #2

Dimension	Scale	Comments
(Factual) Correctness	1-3	Meaning of each score ● 1: This radiology report is mostly incorrect ● 2: This radiology report is mostly correct but important details are missed / incorrect ● 3: This radiology report is correct
Granularity	1-3	● 1: Not enough details ● 2: Includes some details but miss important ones ● 3: Very specific report
Importance/Saliency	1-3	● 1: Ignores the most important part (eg heart is enlarged but the report focuses on the ETT) ● 2: Mostly good, but didnt focus enough ● 3: Really gets at the most important / obvious finding to communicate
Overall how much joy it sparks or Overall rating	1-7	If a radiology student wrote this, what overall numerical grade would you give them?

Proposal #3: (synthesis of radiologist feedback)

Dimension	Scale	Comments
(Factual) Correctness of Stated Claims	1-5	Meaning of each score ● 1: This radiology report is mostly incorrect ● 3: This radiology report is mostly correct but important details are missed / incorrect ● 5: This radiology report is correct
Comprehensiveness	1-5	● 1: Ignores the most important part (eg heart is enlarged but the report focuses on the ETT) ● 3: Mostly good, but did not focus enough ● 5: Really gets at the most important / obvious finding to communicate
Medical Turing Test (we can rename this)	1-5	● 1: The report legitimately does not make sense. Perhaps the author is a computer or otherwise struggles with English. ● 3: The report seems like it was written by someone who can communicate but no medical training (e.g. does weird phrases or is divorced from a standard order of heart/lungs/etc). ● 5: A medical expert wrote this report. It is stylistically good.
Overall how much joy it sparks or Overall rating	1-7	If a radiology student wrote this, what overall numerical grade would you give them?

Conducting the Annotation Task

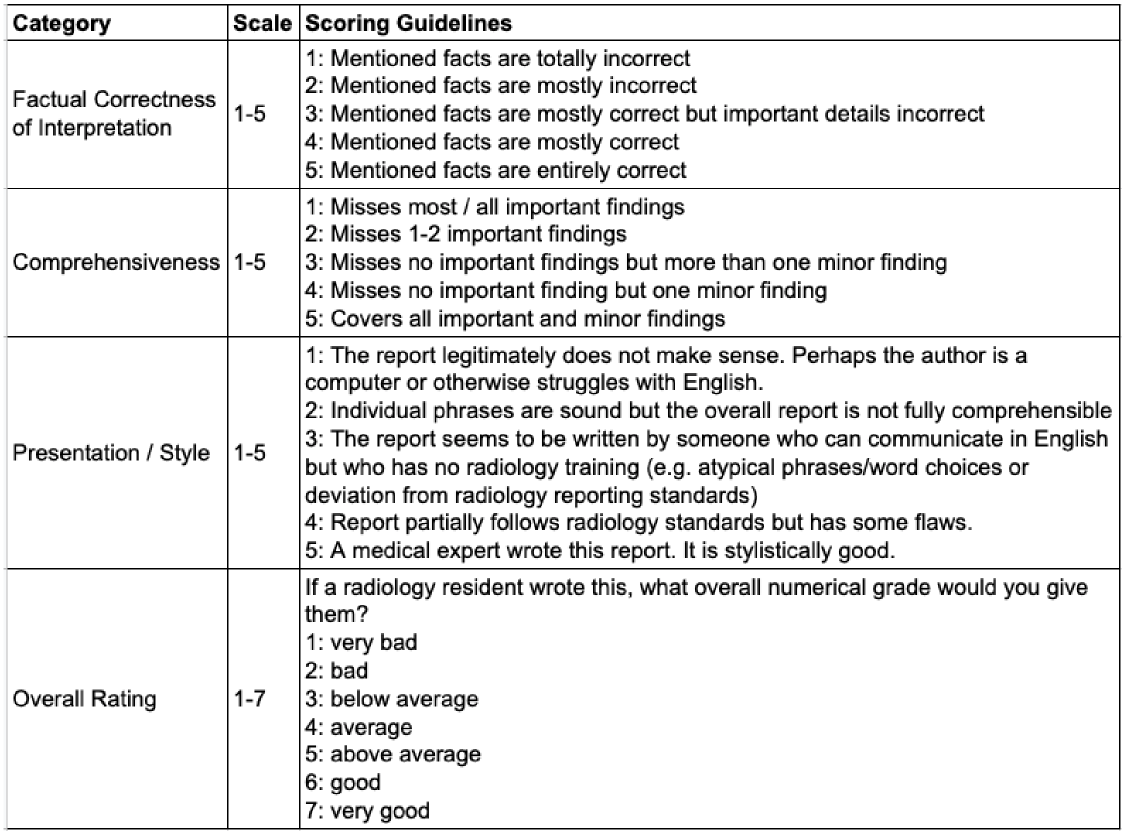
After iterating on possible criteria to see which aspects were relevant when comparing why radiologists preferred one report over another one, we finally arrived at the criteria described in Figure 3-6. One of the main distinctions we ultimately made was to separately measure the analogs of false positives (i.e., were the stated claims correct?) and false negatives (i.e., did you mention all the things you were supposed to?). These concepts correspond to "factual correctness" and "comprehensiveness," respectively. Finally, we look at "presentation / style" because reports can be correct but poorly written (or vice versa). We also include a more traditional "overall" assessment which tries to measure the more standard, generic goodness that collapses all of the different forms of quality into one single dimension.

We incorporated generated captions alongside the reference caption for each of our 167 images. Images and candidate reports were selected so that no candidates (including the reference) contained references to previous exams. We use stronger generation techniques than last time, and further instead of using the same 4 methods for all images, we randomly select the reference and 3 of the following for an image:

- Reference Report
- Nearest Neighbor from training set [35]
- Random Report from training set [35]
- Show, Attend, and Tell [312] with beam size 5
- TieNet [300] with beam size 1
- TieNet [300] with beam size 5

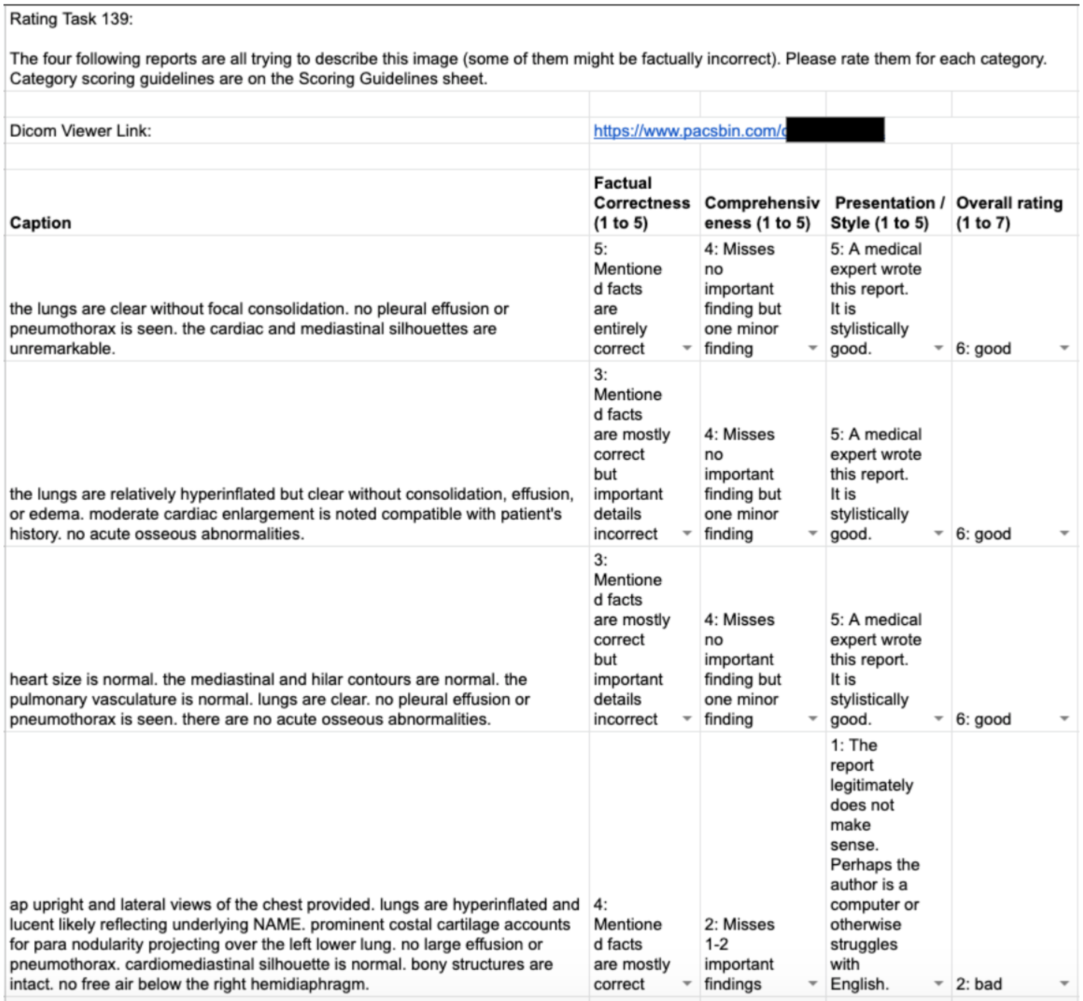
The final annotation task is depicted in Figure 3-7. Annotators examined 167 selected images, each containing four captions. The annotator scored each caption for each of the four criteria based on the scoring guidelines in Figure 3-6. The image was viewable to the radiologist using the pacsbin software, which allowed them to use the PACS format to view the image (i.e., can zoom in, adjust contrast, etc). To address the concern of inconsistency, the annotator was shown all 4 reports for an image at the same time, allowing them to contextualize whether a report is really bad

Figure 3-6: Scoring Guidelines provided to radiologists.



Category	Scale	Scoring Guidelines
Factual Correctness of Interpretation	1-5	1: Mentioned facts are totally incorrect 2: Mentioned facts are mostly incorrect 3: Mentioned facts are mostly correct but important details incorrect 4: Mentioned facts are mostly correct 5: Mentioned facts are entirely correct
Comprehensiveness	1-5	1: Misses most / all important findings 2: Misses 1-2 important findings 3: Misses no important findings but more than one minor finding 4: Misses no important finding but one minor finding 5: Covers all important and minor findings
Presentation / Style	1-5	1: The report legitimately does not make sense. Perhaps the author is a computer or otherwise struggles with English. 2: Individual phrases are sound but the overall report is not fully comprehensible 3: The report seems to be written by someone who can communicate in English but who has no radiology training (e.g. atypical phrases/word choices or deviation from radiology reporting standards) 4: Report partially follows radiology standards but has some flaws. 5: A medical expert wrote this report. It is stylistically good.
Overall Rating	1-7	If a radiology resident wrote this, what overall numerical grade would you give them? 1: very bad 2: bad 3: below average 4: average 5: above average 6: good 7: very good

Figure 3-7: Task presented to radiologists, where all captions are shown at the same time.



Rating Task 139:

The four following reports are all trying to describe this image (some of them might be factually incorrect). Please rate them for each category. Category scoring guidelines are on the Scoring Guidelines sheet.

Dicom Viewer Link: https://www.pacsbin.com/c[redacted]

Caption	Factual Correctness (1 to 5)	Comprehensiveness (1 to 5)	Presentation / Style (1 to 5)	Overall rating (1 to 7)
the lungs are clear without focal consolidation. no pleural effusion or pneumothorax is seen. the cardiac and mediastinal silhouettes are unremarkable.	5: Mentioned facts are entirely correct	4: Misses no important finding but one minor finding	5: A medical expert wrote this report. It is stylistically good.	6: good
the lungs are relatively hyperinflated but clear without consolidation, effusion, or edema. moderate cardiac enlargement is noted compatible with patient's history. no acute osseous abnormalities.	3: Mentioned facts are mostly correct but important details incorrect	4: Misses no important finding but one minor finding	5: A medical expert wrote this report. It is stylistically good.	6: good
heart size is normal. the mediastinal and hilar contours are normal. the pulmonary vasculature is normal. lungs are clear. no pleural effusion or pneumothorax is seen. there are no acute osseous abnormalities.	3: Mentioned facts are mostly correct but important details incorrect	4: Misses no important finding but one minor finding	5: A medical expert wrote this report. It is stylistically good.	6: good
ap upright and lateral views of the chest provided. lungs are hyperinflated and lucent likely reflecting underlying NAME. prominent costal cartilage accounts for para nodularity projecting over the left lower lung. no large effusion or pneumothorax. cardiomediastinal silhouette is normal. bony structures are intact. no free air below the right hemidiaphragm.	4: Mentioned facts are mostly correct	2: Misses 1-2 important findings	1: The report legitimately does not make sense. Perhaps the author is a computer or otherwise struggles with English.	2: bad

or if they merely have nitpicks. We had 3 radiologists complete the scoring task.

3.3.2 Results

For a given image, the annotations allow us to compare the quality of four different reports across four evaluation criteria according to three radiologists. We assess the hypothesis that different dimensions are able to effectively capture different strengths and weaknesses of a report.

Table 3.6: Correlations between dimensions. Each image had 4 dimensions, each dimension's 3 annotator scores were averaged into one score per dimension.

	Correctness	Comprehensiveness	Presentation	Overall
Correctness	1.00	0.89	0.81	0.92
Comprehensiveness	0.89	1.00	0.76	0.92
Presentation	0.81	0.76	1.00	0.82
Overall	0.92	0.92	0.82	1.00

Table 3.7: Correlations between dimensions, separated by annotator. "FC" stands for factual correctness, "C" stands for comprehensiveness, "P/S" stands for presentation / style, and "O" stands for overall.

Annotator 1 (all images)				
	FC	C	P/S	O
FC	1.00	0.85	0.82	0.84
C	0.85	1.00	0.78	0.89
P/S	0.82	0.78	1.00	0.83
O	0.84	0.89	0.83	1.00

Annotator 2 (all images)				
	FC	C	P/S	O
FC	1.00	0.92	0.53	0.91
C	0.92	1.00	0.49	0.93
P/S	0.53	0.49	1.00	0.56
O	0.91	0.93	0.56	1.00

Annotator 3 (all images)				
	FC	C	P/S	O
FC	1.00	0.58	0.41	0.73
C	0.58	1.00	0.37	0.72
P/S	0.41	0.37	1.00	0.43
O	0.73	0.72	0.43	1.00

The Independent Dimension Hypothesis

The hypothesis of this work is that asking annotators about multiple dimensions at once could allow them to disaggregate why a report is good or bad. Ideally, it would find not just good vs. bad reports, but instead allow for a more nuanced analysis where a report might be factually correct in its stated claims without comprehensively addressing all of the abnormal findings in the image. Table 3.6 shows the correlations between the annotators' scores for each dimension. I had expected that these dimensions could be decently independent of each other (i.e., good style shouldn't necessarily indicate good correctness). However, all of the dimensions have a surprisingly high correlation with each other. The pairwise correlations between the three non-overall dimensions are 0.89, 0.81, and 0.76.

In order to understand this phenomenon more closely, we qualitatively investigate the scores assigned to a few of the images by each annotator. Table 3.7 presents the dimension correlations for each annotator. Annotators 1 and 2 had significantly higher dimension correlations than Annotator 3 did. In retrospect, this can be attributed to Annotator 3 being the sole radiologist with whom we iteratively defined the criteria. Even within Annotators 1 and 2, there was a disagreement where Annotator 1 rated all four criteria strongly correlated with each other (six pairwise comparisons all between 0.78–0.89) whereas Annotator 2 felt that correctness, comprehensiveness, and overall were even more strongly correlated (0.92, 0.91, and 0.93) and that all three were much more different from presentation (0.53, 0.49, 0.56).

In an after-the-fact discussion with Annotator 3, they reiterated their prior concern that the distinctions between the dimensions might be too unintuitive. This suggests problems in the task's design, communication, or both. Not only did the other two annotators not understand the intention of trying to measure decoupled dimensions, but even the interviewed radiologist, as it turns out, had a different conception than we did of what "Factual Correctness' and "Comprehensiveness" mean. Per the descriptions from Figure 3-6, we envisioned correctness and comprehensiveness were effectively analogous to precision (i.e., "of stated claims, how many are correct?") and recall (i.e., "did you identify all of the important things you were supposed to?"). However, even Annotator 3 didn't follow the guidelines exactly as written; they'd thought of "Factual Correctness" as something of a catch-all to penalize reports both for incorrect stated claims and missing important claims that felt obvious based on the image. On the other hand, they thought of "Comprehensiveness" as refering to the structure of the report (i.e., "does it talk about the heart? the lungs? support devices? etc.") instead of thinking about comprehensiveness of the salient/active observations.

3.4 Project "Evidence"

The original goal of the project was to evaluate the quality of generated reports. Ideally, I hoped to collect expert-annotated data and develop a new clinically-informed metric which agrees with the annotators better than previous metrics do. However, neither data collection resulted in satisfyingly high quality data.

Some effective altruists engage in the practice of self-critique. GiveWell[4] and

[4] https://www.givewell.org/about/our-mistakes

80,000 Hours[5] each maintain a public directory of mistakes they have made and what they've learned. Although some psychologists argue that self-critique can be counterproductive because it elicits demobilizing reactions [180], others argue that it can be part of the path to increased self-awareness and associated with many benefits [79]. Julia Galef – a member of the Rationalist community, which has overlap in methods and membership with Effective Altruism – recommends that individuals and institutions should adopt the "scout mindset," seeing the world as it is rather than as we want it to be [92]. In particular, she argues that a mindset focused on learning and improvement is able to adapt to adversity more readily than one whose identity often feels threatened from challenges to core beliefs. Throughout this section, I demonstrate the constructive value of self-critiques to identify areas for improvement.

In the early 1980s, Boeing developed the "Plus-Delta" reflection tool (also known as "Do Again / Do Better") to identify both bright spots and challenges at the end of meetings and projects [189]. This can be a helpful tool both because it identifies actionable information and also because it pairs constructive criticism with compliments, which might allow responders to feel more comfortable giving feedback without being seen as negative. For each annotation experiment, I do a Plus-Delta reflection, and then I summarize lessons learned.

3.4.1 Self-Critiques of Attempt 1

For the first round of annotations, we conducted an annotation task to for two radiologists to rank 400 candidate reports from 100 radiology images.

Plus (Bright Spots)

Not Overselling Data Quality. One successful effort from this project is that the radiologist annotations were not reflexively reported as the unimpeachable truth. There have been instances of human annotations which later works have demonstrated

[5] `https://80000hours.org/about/credibility/evaluations/mistakes`

inconsistencies or human error [32], but this work did not fall into that framing. Instead, upon identifying that the judgments had inconsistencies and design limitations, the effort was framed as a learning opportunity / pilot study instead of glossing over those challenges and marketing the dataset as useful for any researchers doing their own evaluations of generated reports.

Collaboration Tools. To that end, we did a good job creating the tools for collaborate with radiologists, particularly the "1-pager" shown in Figure 3-2. This document helped both with identifying radiologists interested in talking (i.e., circulate the 1-pager to our network and see who responds) and also with orienting the clinicians towards the kinds of trends we would like to study. These conversations helped identify areas of consensus amongst the three radiologists: correctness as a stated priority, a move towards regular structure in the field, and that the PNG format of images (instead of DICOM) was fine for this task. One radiologist mentioned the importance of having images which contain a diversity of diagnoses.

New Annotation Task Framing. Based on the pre-annotation interviews, we also saw successes in designing the annotation task: using a ranking-based approach allowed for annotators to provide judgment more consistently than trying to assign an objective, overall "goodness' score to candidate reports in a vacuum. Particularly, this format encouraged us to display multiple candidates at once, thus priming annotators to rate reports as good/bad in context, depending on whether there will small nitpicks between two good candidates or whether one report is obviously worse than another.

Annotator Explanations and Confidence Scores. Finally, it was useful to ask annotators for explanations and to estimate their confidence. This allows researchers to both sanity check the annotations and to learn what factors radiologists seem to identify as important to their process. By asking whether they felt confident that a fellow radiologist would correctly identify the best candidate, we were able to measure how replicable they felt their "easy" judgment was without worrying whether lower quality models would have reports that were ranked slightly differently. The wording of this question was well-suited for this task.

Delta (Opportunities to do Better)

Weak Generative Models. One shortcoming is that the caption generation methods were very simple. Although a subset of trigram, random report, and nearest neighbor could serve as useful baselines, there are no deep learning models generating plausible reports. As shown by the annotators, reference reports were rated the best, and no other reports were even close. This does allow for measuring the quality of reports in broad strokes, but it does not allow collecting annotations that would allow for fine-grained differentiation between two high-quality candidates.

Reports Not Following "Image Captioning" Framing. Another limitation encountered in this work is that many of the reports in the dataset were not fully grade-able, because they did not capture everything a radiologist would have available to them when really performing their work. Over half of all reports in the annotation pilot referenced previous radiographs, which meant annotators needed to make their best guesses about unseen data. Additionally, although we only show the frontal chest x-ray, reports are usually written using multiple views, such as frontal and lateral.

Information Inefficiency of Ranking-based Task. This task adopted a rating-based approach, however the comparisons did not offer enough information about why a given report was good or bad. Some lowest-rated reports were egregiously low-quality whereas other ones may have been adequate but unfortunately faced even better competition. Simply using a ranking-based approach does not allow for annotators to signal whether a report is really well-done or completely incomprehensible. Maybe with thousands of such reports that might be fine, but data is more expensive to collect from expert annotations; one radiologist estimated that annotating 200 studies may fetch as much as $5,000 if done for a pharma study.

Lessons Learned

Summarizing the positives and negatives from this experiment, a followup study (which we ultimately did perform, as seen in Section 3.3) would build upon this work by focusing on:

- Continuing to use collaboration tools such as interviews and the 1-pager in Figure 3-2.
- Generating candidate reports using better language models.
- Filtering the experiment's reports to be information self-contained within the x-ray image (i.e., no reports which reference previous scans).
- Selecting a different annotation task framing to allow annotators to convey more information than mere ranking allows.

3.4.2 Self-Critiques of Attempt 2

For the second experiment, we collected annotations for 501 reports from 167 images using. We used strong baselines and deep learning methods to generate these reports. The reports were evaluated using multiple dimensions of quality: factual correctness, comprehensiveness, presentation / style, and overall quality. Although this work was informed by lessons learned from the prior annotation study, it also suffered from shortcomings to learn from.

Plus (Bright Spots)

Learned from Previous Mistakes. The first thing this study did was learn from the Plus-Delta of the previous experiment. This can be seen with:

- Better generative models (TieNet and Show-Attend-and-Tell instead of 3-gram) for annotators to contrast multiple plausible reports.
- Improved annotation task, where annotators can combine the benefits of both ranking (comparisons to help contextualize differences) and rating (express whether a report is slightly worse or much worse than another report) by giving 1-5 scores to all four candidate reports at once. Additionally, we did not use the same four generative models for every image, but instead we mixed the use of models just in case there had been a bias introduced by a particularly bad generative method.
- Continued using collaborative tools for the design of the task with radiologists.

Using the worksheet in Figure 3-4, we were able to iterate on which dimensions could be a meaningful way to break down how to assess the quality of a report.

Designing Better Annotator Experience. Seeking the feedback from radiologist collaborators allowed us to identify ways to improve the experience of the volunteer annotators. For instance, the radiologist who served as an annotator for both experiments suggested that we should try using a DICOM viewer (which allows for zooming and changing the contrast) after all for the second attempt. Additionally, they identified that the previous submission form (google form) forced the annotators to do the task in batches of 25, which was difficult to fit into their day. By switching to a google sheet (with links to the images), annotators were able to make partial progress more easily.

Strong Baselines. We find that neural network approaches offer the strongest performance on these models, however 1-NN methods are robust contenders, particularly with regards to clinical efficacy as measured by CheXpert predicted label F1 scores. In light of this — and the ease with which one can implement nearest neighbor methods — I recommend including a 1-NN baseline for future report generation projects as a new best practice. This will help disentangle the benefits of better image encoding vs. better feature-to-text decoding. This is especially relevant because we expect that with transformer-based language modeling, it is likely that feature-to-text decoding will improve in the near future.

Delta (Opportunities to do Better)

Miscommunications with Radiologists. Despite our attempts to integrate clinical collaborators into the annotation design task, we nonetheless ran into serious miscommunication challenges. To varying degrees, all three annotators misunderstood the criteria for the different dimensions. For the two radiologists that were not able to attend design meetings and only participated in reading+annotating, they had strong correlations across all dimensions (thus defeating the experiment's attempt to measure meaningfully different aspects of quality). However, even the radiologist who did participate did not annotate as we'd anticipated: their understanding of correctness and comprehensiveness was different from our understanding. This likely indicates that we did not present the definitions in an accessible way; the wall of text explaining the definitions was likely scanned quickly, similar to how some participants quickly scan over terms of service forms for software.

Lack of Diversity in Diagnosis. One issue in retrospect is that more x-ray images were normal and without any medical issues than originally realized. This was a programming bug in interpreting CheXpert's different labels (1, 0, -1, and u). Because CheXpert evaluates for 14 diagnoses, there are 4^{14} unique CheXpert profiles. As a result, merely checking for a unique CheXpert output profile was insufficient. For instance, the following two reports would register as having different CheXpert output profiles:

- "lung volumes are normal. there is no central vascular congestion or overt pulmonary edema. mediastinal and hilar contours are normal. heart size is normal."
- "ap and lateral chest radiographs were obtained. the lungs are well expanded and clear. there is no focal consolidation, effusion, or pneumothorax. there is no free air under the diaphragm. gastric distention is better appreciated on the abdomenal radiograph."

Although they both essentially say "the patient is fine and nothing is abnormal," they have different explicitly negative conditions. The first report only describes a few CheXpert categories which are affirmately not present (e.g., enlarged heart, edema). On the other hand, the second report explicitly states there is no consolidation, effusion, or pneumothorax. When conditions are explicitly mentioned, CheXpert labels them with 0 (negative) whereas when they are unspecified, it is labeled with u (uncertain) despite implicitly being a negative indication. A majority of studies were not abnormal, which on the one hand is representative of the MIMIC-CXR database but on the other hand has diminishing returns in how valuable it is to rate how many ways a report can adequately say "the patient does not have problems." Future studies should select for a diversity of images and manually inspect them before the full-scale annotation task. The domain expertise barrier intimidated and discouraged members of the computer science team from feeling able to notice this problem before the annotation task began.

Lessons Learned

Multiple issues, including both the lack of disease diversity and the confusion in dimension definitions could have been mitigated by having each annotator do a 10-20 image mini-pilot / sanity check to identify any such hiccups. A common barrier for these pilots can be wanting to batch the requests to annotators, especially if they are volunteering their time. These concerns could be mitigated through upfront communication, workload expectations, and (ideally) funds to compensate annotators for their time [250]. At the very least, we should have spoken with each annotator to discuss the evaluation guidelines instead of merely including it on the cover page that they likely did not inspect very closely.

Chapter 4

Racial Disparities in End-of-Life Care

In this chapter, I describe a case study where I extend prior work by deploying a model to address racial disparities in end-of-life care. Although the specifics around the deployment are proprietary, I discuss a similar kind of model trained on public data. The contribution of this work is not one of technical innovation. Instead, this work demonstrates the importance of framing a problem actionably and working to deploy it.

4.1 Disparate Treatments During End-of-Life

Previous studies have identified that different patients experience end-of-life (EOL) differently [190, 158, 109]. Researchers looked at "invasive" care (i.e., interventions that are unpleasant, such as when a tube is inserted into the patient's throat to try to prolong life) vs. comfort-based case (i.e., hospice). The main finding was that white patients received smaller amounts of invasive care than nonwhite (in particular African American and Hispanic) patients. Renowned author and surgeon Atul Gawande has found that patients who are empowered to make informed EOL decisions overwhelmingly choose to live their final months "with dignity" at home instead of pursuing overly medicalized procedures cooped up in a hospital [97]. So why did these studies find that white patients seemed to be transitioning to hospice care earlier and at higher rates?

Hanchate et al. [109] suggest that “this may stem from distrust of the medical care system or from economic constraints.” When further invasive procedures are unlikely to return a patient to a normal lifestyle, their doctor may recommend withdrawing treatment and transitioning to comfort-based measures to ensure the patient does not suffer. However, if the patient (or healthcare proxy) doesn’t trust the doctor to really be acting in their interests, then it could lead them to question the assessment (e.g., maybe the hospital doesn’t want to use resources), and instead demand additional invasive interventions [95]. Poor trust has specifically been shown to impact end-of-life care; family members of African American patients are more likely to cite absent or problematic communication with physicians about EOL care [114].

To better understand and contextualize where some Black patients’ distrust could be coming from, we can look to Harriet Washington’s 2007 book *Medical Apartheid*, which argues that the exploitation of African Americans by medical institutions throughout American history has created “Black Iatrophobia” [301]. Most readers will likely be familiar with one of the most high-profile examples: the Tuskegee Syphilis Study, where a group of African American men with syphilis were denied treatment for three decades because doctors wanted to study the progression of the disease [52]. Washington contends this was not an isolated issue but is the most infamous example of a broader pattern where experimentation happens on the marginalized populations which are least able to fight back. Going back to 1801, Thomas Jefferson injected 80 of his own slaves with smallpox to prototype vaccines [130]. In the late 1840’s, Dr. James Marion Sims (considered by some to be "the Father of Gynecology") surgically experimented on and mutilated his female slaves without anesthesia [160].

In my 2018 Master’s Thesis [30], I analyzed the role that mistrust can play in these disparate EOL treatments/outcomes. I found that statistically-derived proxies for mistrust (e.g., likelihood for doctors to mention “noncompliance” with recommendations or instructions) were even more indicative of EOL disparities than race. The features most associated with high levels of mistrust scores were about how the patients interacted with the care team: agitated, in pain, and sometimes physically restrained. Additionally, I found higher levels of mistrust in the Black patient population than the white patient population [29, 30].

Racial disparities caused by differences in treatment and institutional access reflect a systemic bias that should be addressed. But how could one try to act upon insights derived from a research study? There are multiple interventions to address these disparities. One approach could be interventions such as training and recommendations to the hospital care staff about how to recognize and mitigate the biases that arise. However, this could be difficult because if the patient has had negative experiences with medical institutions for decades, it is unlikely that any training would be able to equip medical strangers with enough ability to earn the patient's trust [301]. Advance care planning (ACP) tools allow a patient to express their EOL preferences after ample time for reflection, though white patients have higher rates of ACP usage than Black and Hispanic patients [266, 135, 267]. A preventative solution could be to try to equalize ACP rates between races before any of them arrive at the hospital. This would allow patients to decide their own treatment plan without the pressures and fears of making the "wrong" decision in-the-moment.

In Section 4.2, I describe a model, trained on public data, which shows a proof of concept for how to build a mortality risk stratification model on clinical data. In Section 4.3, I describe how deploying this model for preventative care can identify more patients which would benefit from pre-hospital advance care planning.

4.2 Building ML: EOL Risk Prediction

Advance directives – such as living wills and designated healthcare proxies – are written, legal instructions about what actions should be taken for one's health if they are unable to make decisions for themself. Because "[u]nexpected end-of-life situations can happen at any age," the Mayo Clinic recommends that "it's important for all adults to prepare these documents" [272]. However, only 1-in-3 American adults have completed an advance directive [314]. In an ideal world, every patient would have a conversation with their doctor and family about advance care planning. However, with limited resources, it may be necessary to identify the patients at highest near-

Table 4.1: Rates of patients who pass away within one year of discharge rates by race. This uses the MIMIC cohort of patients with a recorded post-discharge death date.

race	**N**	**# positive**	**% positive**
White	6654	3492	52.5
Not Specified	1086	420	38.7
Black	602	370	61.5
Other	320	171	53.4
Hispanic	177	81	45.8
Asian	153	104	68.0
Total	9144	4699	51.4

term mortality risk for intervention. If anyone should have their wishes codified and their papers in order, it should be them. To address this, there have been efforts to build machine learning models to predict which patients are at highest risk for near-term mortality [13].

In this section, I demonstrate a simple machine learning model for 1-year mortality risk stratification. This model is trained on public EHR data, and has not been deployed. The code for this model is publicly available.[1]

4.2.1 Data

I use MIMIC III, a publicly available dataset of ICU stays [132]. The task is a binary prediction of whether a discharged patient will pass away within 12 months of discharge.

Because of a data collection bias detailed in the bias audit described in Section 4.4.1, we only predict for patients that we know passed away at some point post-discharge; patients reportedly still alive are filtered out. We filter out patients who have a code status of "comfort measures only," have an ICD code for palliative care, are under 18 years old, or are discharged to hospice. The cohort contains 9,144 patients. Table 4.1 shows the breakdown in 1-year mortality rates by race.

Features for the model are extracted from the EHR, including demographics (race, gender, age), risk score (SOFA, OASIS, SAPS II), Elixhauser Comorbidities [77],

[1] `https://github.com/wboag/eol-mort-pred`

Table 4.2: Model performance by race, evaluated on the cohort from Table 4.1. We report the mean across 20 runs, plus or minus the standard deviation.

race	N	AUC	Precision	Recall
White	6654	0.717 ± 0.009	0.832 ± 0.025	0.165 ± 0.007
Not Specified	1086	0.697 ± 0.024	0.857 ± 0.093	0.100 ± 0.023
Black	602	0.674 ± 0.034	0.865 ± 0.067	0.200 ± 0.046
Other	320	0.729 ± 0.036	0.906 ± 0.088	0.172 ± 0.059
Hispanic	177	0.681 ± 0.041	0.747 ± 0.239	0.166 ± 0.086
Asian	153	0.644 ± 0.082	0.841 ± 0.119	0.281 ± 0.068
Total	9144	0.716 ± 0.008	0.837 ± 0.020	0.165 ± 0.004

final code status, and admission metadata (elective/emergency, admission location, discharge location, insurance, religion, marital status).

4.2.2 Methods

The data is split 70/30 into train and test sets. The top 10% highest risk patients in the test set are predicted as “will die within a year.” The AUC, precision, and recall of models are reported from averaging across 20 randomly generated train/test splits.

I employ a gradient boosting model using the XGBoost software package.[2]

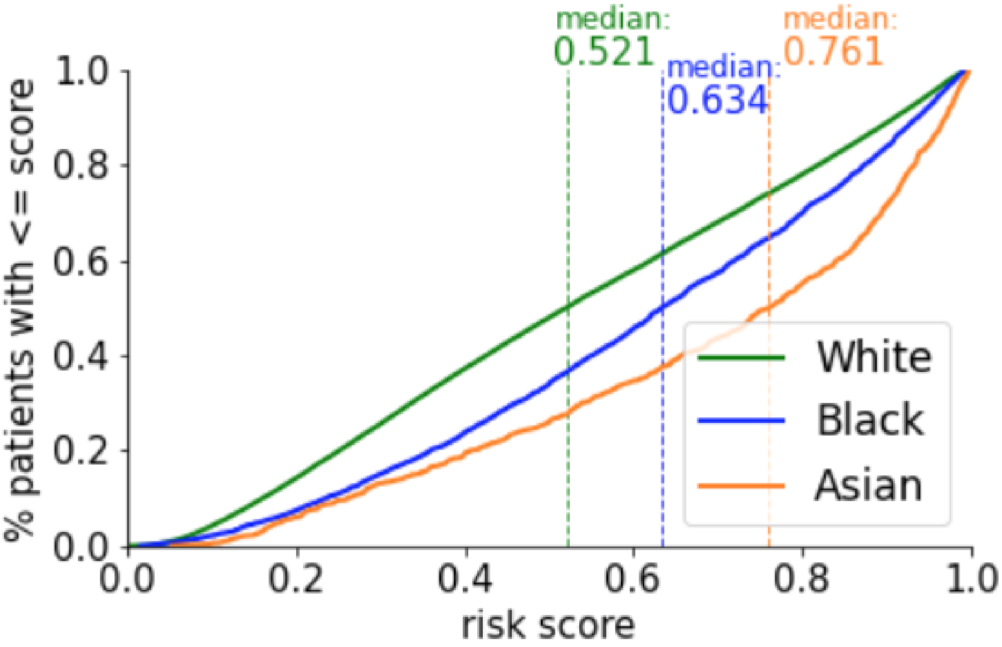
4.2.3 Results

Table 4.2 lists the performance of the model when evaluated on just the patients within each racial group. One overall trend is that the average is largely determined by performance on the white patients, which represents nearly 75% of the dataset. Additionally, the more patients in a racial group, the better the AUC for that group tends to be, because the training process optimizes the model to do well on the most commonly occurring patterns. Finally, we can see that groups with the highest recalls (i.e., the groups which benefit most from this intervention), Black patients and Asian patients, are exactly the ones with the highest mortality rates in Table 4.1.

Figure 4-1 shows the distribution of risk scores for Black, Asian, and white pa-

[2] https://xgboost.ai

Figure 4-1: Distribution of risk scores for white, Black, and Asian patients. The dotted line indicates the median risk score for that group.



tients; as expected, the groups with higher risk scores are the same groups with higher recalls. The property of an intervention which structurally targets the areas in highest need is analogous to what political scientists call a “thermostatic model” [269]. Not all prediction models have this property. For instance, using a sparser and higher-dimensional set of features (just demographics and bag-of-ICD codes) achieves very similar AUCs but sees Black and white patients achieving virtually identical recall scores, despite having different baseline rates of 1-year mortality.

From an inspection of which features are most informative, we find that SOFA score, presence of the metastatic cancer Elixhauser comorbidity, being discharged to a short-term hospital, and a code status of “dncpr” (i.e., if patient cannot breathe, “Do Not attempt CPR”) are most associated with 1-year mortality. On the other hand, the following were most associated with 1-year survival: a code status of “full code,” (i.e. if patient cannot breathe, do everything you can to save them) having a life partner, being discharged home, and having an elective admission in the first place.

4.3 Upstream Intervention: ML for ACP

Academic papers describing models like the one above don't directly help patients. In Chapter 2, I examine the impact of an academic paper; its goal was not to save lives itself, but instead, the work helped shape design choices for later researchers' followup investigations. In this section, I discuss one pathway for translating these tools into the clinic to improve outcomes for patients. Because the actual work itself was proprietary, I do not report on what the specific, measurable outcomes that have been collected since deployment. However, I discuss the approach and why it was very well-suited to address the issues identified in prior work.

As suggested by [109] and explored further in my Masters thesis, one barrier to health equity in EOL is mistrust in the doctor-patient relationship. And even after a staff training or two, it is unlikely that the hospital care team will be able to earn the patient's trust in a high pressure environment. Instead, I propose to intervene upstream by working with primary care physicians (PCPs) who patients already have an existing relationship with.

After 2018, I worked with a healthcare company to deploy a model which identified patients at high risk for 12-month mortality. Patients most at risk were flagged in the population health app used by their primary care doctors partnering with the organization. This addressed a large problem: not only is the PCP more trusted than unfamiliar hospital workers, but this solution also tackles the problem while the patient is still conscious and able to decide for themselves whatever it is they want to do.

The proof of concept model in Section 4.2 makes predictions for patients discharged from the hospital, but in practice that tool would not be as useful. Patients who spend time as inpatients are already interacting with the healthcare system, and especially if they have been to the ICU, they may have already had an ACP discussion. On the other hand, by partnering with a primary care doctor, a model could run on observational data for all patients in their practice. Further, it would have access not just to inpatient data but also other information such as outpatient visits and perhaps medication refills. By predicting for the whole population of patients, there is the opportunity to engage in outreach to high-risk patients that have had little contact with healthcare systems who might otherwise "slip through the cracks."

One shortcoming of this approach is that by working with primary care doctors and through their network, we are not able to identify any patients who don't have a doctor (either because they don't have health insurance or because they are strongly disconnected from the healthcare system). In principle, public health organizations should be able to apply similar kinds of strategies to identify and treat those hardest-to-reach patients. In Rockford, Illinois, an interdisciplinary team of social workers, firefighters, nonprofit staffers, etc., was able to eliminate chronic homelessness [116]. The team cited data as one of the ingredients, though the most important part was not technical; the essential element was uniting the right people and engaging in productive communication across the agencies.

4.4 Project "Evidence"

For this case study, I demonstrate the importance of algorithmic audits for the models that we build. Of course audits are not all that must be done, but they is an important first step in responsible machine learning.

4.4.1 Algorithmic Audit

Algorithmic audits have been one of the most impactful components of AI Ethics research and journalism in the last decade. In 2016, ProPublica's "Machine Bias" investigation of racial biases in crime prediction [11] attracted immense attention to the field of algorithmic bias and AI ethics. The influential FAccT paper to date is "Gender Shades" (2415 citations vs. 574 for the second-most-cited FAccT paper), which focuses on the intersectional (by gender and skin color) biases in facial recognition tools [43]. Additional high-profile works also audit machine learning models: Bender et al. [22] critique the environmental and financial costs of training large language models, Ribeiro et al. [242] analyze over 330,000 videos to better understand how youtube's recommendation algorithm enhances radicalization, Raghavan et al. [231] examine the claims and practices of companies offering algorithms for employment assessment, and Obermeyer and Mullainathan [210] find racial biases in the labels being used to guide health decisions for 70 million people. In a systematic review of algorithmic accountability, Wieringa [303] identifies 93 core articles which model a range of methods in algorithmic accountability across actors, forums, relationships, content, and consequences. Senators Booker and Wyden have even introduced the Algorithmic Accountability Act [37], which would require companies to fix flawed computer algorithms that result in inaccurate, unfair, biased decisions.

Enacting a core principle of Data Feminism, algorithmic audits challenge power by scrutinizing the tools being used for high-stakes decision making. When analyzing this task, I found that the model was outputting surprising results. When I interrogated the reason further, I identified a bias in the labels, which — similar to Obermeyer and Mullainathan [210] — mistakenly recommended over-allocating resources for white patients. In this section, I demonstrate how an algorithmic audit provides actionable opportunities to identify and address biases that arise.

I demonstrate the audit on a cohort different from the one used in Section 4.2 (i.e., the "filtered" cohort); this audit cohort (i.e., the "unfiltered" cohort) includes *all* (adult, non-CMO) patients discharged from the hospital, not just ones which have a recorded post-discharge death rate. The reason for the different cohorts is explained in further detail below, but essentially: this audit detected a bias in the "unfiltered" cohort, so the "filtered" cohort was constructed during the audit's "mitigation plan" step.

I use the audit methodology from Raji et al. [234], which proposes the SMACTR (Scoping, Mapping, Artifact Collection, Testing and Reflection) framework. I apply each stage of the framework to the model trained on the "unfiltered" cohort.

Scoping

.

In this stage, we specify the requirements and expectations of the product or feature. Raji et al. [234] suggest creating *Ethical Review of System Use Case* or *Social Impact Assessment* documents. This stage begins to anticipate potential use cases, motivations, intended impact, and risks.

The predictive model attempts to satisfy multiple goals from different stakeholders. If everything is working as intended, the patients at highest risk for near-term mortality are able to have an ACP conversation with their PCP that they trust. This process aims to address the disparity in advance care planning rates between white and nonwhite patients, and pursues this by working through the existing relationship the patient has with their PCP. Two potential harms that might arise include:

- **Algorithmic Bias**: If the model amplifies existing biases (e.g., by race, gender, religion), then this tool will not achieve the goal of equity.
- **Losing Trust**: This intervention is done in partnership with PCPs to "level up" the analysis so that it could be used in a meaningful and helpful way to the patient. It would be counterproductive if the tool "leveled down" the relationship between the doctor and patient, eroding the existing relationship.

The risk of algorithmic bias is nontrivial, and will merit investigation for common pitfalls during the "Testing" stage.

The risk of the algorithm undermining the patient-PCP relationship can be mitigated through the interface of deployment and how the PCP approaches the conversation with the patient. Hospice workers understand these are delicate conversations with vulnerable people, and they have a recommended set of principles for bringing the topic up [97]; the PCP will need to devise an appropriate way to initiate the conversation. Likely, they would get the risk alert, and tailor it to that patient (e.g., "Sarah, I noticed you're 73 and are still living independently, but we don't have any records on file about what your priorities are and what sacrifices would you be willing to make, or not make, to maintain those priorities.") as opposed to saying something distant and alarming like "the AI said you are at high risk, so we should probably talk about this now."

Table 4.3: 1-year mortality rates by race for the "unfiltered" cohort.

race	**N**	**# positive**	**% positive**
White	22482	3492	15.5
Not Specified	2860	420	14.7
Black	2435	370	15.2
Other	1825	171	9.4
Hispanic	1107	81	7.3
Asian	730	104	14.2
Men	18042	2525	14.0
Women	13412	2113	15.8
Total	31892	4699	14.7

Mapping

.

This is a stage for reviewing what is already in place, including the roles and responsibilities of internal stakeholders and collaborators. Recommended documents for this stage include a *Stakeholder Map* and *Ethnographic Field Study*, which are defined in more detail in Raji et al. [234].

If a model were to be deployed, the most meaningful pre-deployment metric would be recall (on the heldout test data), because it measures 'of all of the patients who should receive the intervention, how many actually do?" Once a model has been adopted into the clinical workflow, it is more important to measure the model's actual impact, for instance by comparing whether ACP rates have changed from the previous year or if the racial EOL treatment gap is closing.

Artifact Collection

.

The artifact collection stage is when the documents from the prior stages are aggregated to prioritize opportunities for testing. The output of this process can take the forms of a *Design Checklist* and producing a *Datasheet for the Dataset* [98] and *Model Card* [184].

Because not every Datasheet question is about the data generating process, I present an abbreviated Datasheet here. One challenge using a public dataset is that many of the questions pertaining to collection methods and intended use don't have natural answers. The documentation for MIMIC can be found here.[3] Answers to a few relevant questions include:

- Q: What data does each instance consist of? "Raw" data or features?
  - demographics (race, gender, age);
  - risk score (SOFA, OASIS, SAPS II) derived from EHR measurements such as body temperature, oxygen saturdation, and more;
  - Elixhauser Comorbidities [77] derived by clustering important ICD-9 codes together;
  - final code status from the EHR; and
  - admission metadata (elective/emergency, admission location, discharge location, insurance, religion, marital status).
- Q: Is there a label or target associated with each instance? If so, please provide a description.
  - Yes, the date of death. Per MIMIC-III documentation,[4] dates of death were obtained from both the EHR (for in-hospital) and social security master death index (for out-of-hospital). This project will make use of the social security-collected data.
- Q: Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset
  - Yes, we have information on patient gender, race, and age. This information is entered into the EHR. Sometimes race is not able to be identified and is entered as "Not Specified." Table 4.3 demonstrates the breakdown by demographic.
- Q: Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was

[3] https://mimic.mit.edu/docs/iii/tables

[4] https://mimic.mit.edu/docs/iii/tables/patients

Table 4.4: Model performance by race, trained on the "unfiltered" cohort. We report the mean across 20 runs, plus or minus the standard deviation.

race	AUC	Precision	Recall
White	0.839 ± 0.004	0.546 ± 0.013	0.377 ± 0.013
Not Specified	0.828 ± 0.016	0.500 ± 0.051	0.319 ± 0.043
Black	0.829 ± 0.019	0.524 ± 0.067	0.341 ± 0.052
Other	0.823 ± 0.026	0.486 ± 0.088	0.285 ± 0.061
Hispanic	0.843 ± 0.035	0.432 ± 0.118	0.211 ± 0.071
Asian	0.871 ± 0.021	0.512 ± 0.064	0.418 ± 0.081
Total	0.839 ± 0.005	0.536 ± 0.012	0.363 ± 0.010

unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

- The MIMIC-III documentation does not specifically mention anything. This is one large shortfall of using a large public dataset; there probably are forms of missing data, but because of how many workers (nurses, doctors, technicians, researchers) contributed to assembling it, it may be too difficult to identify the many ways missingness may manifest.

Testing

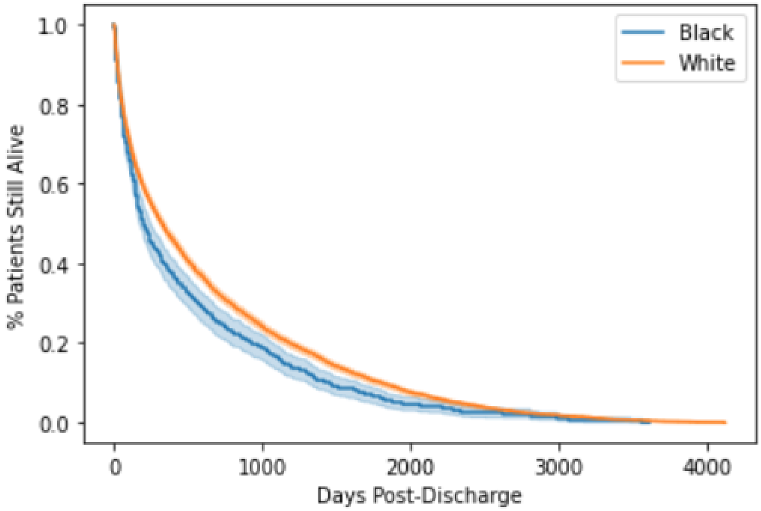
.

In this stage, we execute tests to gauge how well the system complies with the prioritized ethical values from earlier stages.

I fit the model on the data from Table 4.3, and show those results in Table 4.4. Although the model's AUC being better for white patients than Black patients is not necessarily surprising, it **was** unexpected for me to see that the model has a higher recall for white patients than Black patients. Looking back at Table 4.3, we can see that white patients are said to have the highest post-discharge mortality rates, which goes against a large body of work that indicates Black patients have worse health outcomes across a large number of settings, including severity of illness and life expectancy [304].

I inspect this surprising result further in Figure 4-2, which shows a Kaplan Meier

Figure 4-2: Kaplan Meier curve of patients that die post-discharge.



curve of mortality post-discharge. More inline with my expectations: of the patients who do die, Black patients die at a faster rate. Therefore, why does the "unfiltered" cohort indicate that white patients are at higher risk for mortality than Black patients?

Upon closer look, I find this discrepancy is caused by a collection / missing data error.[5] MIMIC-III uses social security records to collect patient mortality information for patients who passed away outside of the hospital. But in 2011, the Social Security administration reinterpreted section 205(r) of the Social Security Act to prohibit the agency from sharing data obtained from state records [198]. The result was "a 40% drop in the capture of deaths." In particular, it seems that Black patient deaths are being disproportionately under-counted, resulting in an under-estimate of their risk.

Reflection

.

In this final stage, we examine the results of the tests and analyze their relationships with expectations clarified in the Scoping section. The essential output of this process includes recommendations for mitigating any harms that have been identified. These recommendations can take the forms of an *Algorithmic Use-related Risk*

[5] `https://github.com/MIT-LCP/mimic-code/issues/1199`

Analysis and FMEA, a *Remediation and Risk Mitigation Plan*, an *Algorithmic Design History File*, or an *Algorithmic Audit Summary Report*, which are defined further in Raji et al. [234].

Another principle in Data Feminism is to consider context [64], including how the data generating process shapes the data we work with. Often, data is presented as if it is fully "cooked," and its assumptions and imperfections are not properly understood by the data scientist. We were able to identify the mechanism through which this bias arose (i.e., SSA 2011 policy change in data sharing practices) because it was discussed on a public MIMIC-III forum.[6] However, there are even more challenging scenarios where the biases are so invisible that one will never find them without talking to the care workers that entered the data themselves. For instance, during a discussion with a nurse from Beth Israel Deaconess Medical Center (the hospital MIMIC is from), I asked about how to understand the coded items in the chartevents table. She described how some of these entries come from an overwhelming number of pop-ups in the EHR which few people would have enough time to comprehensively fill out. Further, she said that data entry can vary from nurse to nurse and although some nurses do code "met with the family" in the structured chart, she has only ever indicated family meetings in her nursing notes. As was discussed earlier, scenarios like this demonstrate the difficulty of using large public datasets because researchers have less access to the participants in the data generating process.

This audit is very similar to Obermeyer and Mullainathan [210] which also found a racial bias in the labels being used to train a model for targeting resources. The label bias in both models led to a systemic under-estimating of Black patient's health needs. A critical part of the responsibility for a tool like that is to ensure the service is mitigating biases rather than exacerbating them.

Specifically to address the identified problem of unreliable labels, I filter the cohort to *only* contain patients who have passed away (some within a year, some not). Doing this ensures we have accurate data about the dates of death for all patients in the data. The filtered cohort is the one in Table 4.1, for which the case study initially

[6] `https://github.com/MIT-LCP/mimic-code/issues/1199`

reports results.

Chapter 5

Tech Worker Collective Action

Of course data scientists have many technical contributions to make to ethical AI, but they are also well-situated to push their influential employers to act more ethically. In this chapter, I shift from computation-centered projects and explore the under-studied landscape of tech worker collective action as a mechanism for ethical computing. One of the seven principles of Data Feminism is to "make labor visible;" in this chapter, I examine that principle both literally and conceptually.

This work was done in collaboration with Bianca Lepe, Harini Suresh, and Catherine D'Ignazio [33].

5.1 Problem Definition

Since Fairness, Accountability, and Transparency (FAccT) launched as a conference in 2018, the community has experienced rapid growth. The 208 FAccT papers published between 2018-2021 feature impactful work, including scholarship that anchored the conversation for facial surveillance bans across the US [43], improved a deployed ML model that was making racially-biased predictions for millions of patients [210], and challenged powerful corporate interests [22]. Other work has built useful toolkits to audit systems for censorship [316], exclusionary design [2], community authorship diversity [58], context-sensitive documentation [183], and more.

In FAccT and beyond, there has been very thoughtful research that offers tangible visions for how tech companies could be applying responsible computing practices, such as a procedure for internal algorithmic audits [235], co-designing checklists for AI fairness with practitioners [176], and a series of questions for self-assessment of whether AI products are respecting human rights [274].

However, most of these approaches require voluntary commitment from relevant corporations. Although corporate buy-in can make ethical computing easier, it is often the case that profit-maximizing organizations resist these efforts precisely where they are most needed.

For instance, Facebook's revenue model comes from selling ad placements to display to its users, which incentivizes the organization to try to maximize user engagement. There has not been a dearth of what responsible social media metrics *could* look like; in 2018, Cortico developed four indicators of conversational health: shared attention, shared reality, variety of opinion, and receptivity [248]. Although Facebook CEO Mark Zuckerburg has claimed that the Facebook algorithm does not optimize for "what [users] click on or will make us the most revenue," but rather "what people actually find meaningful and valuable" [148], this characterization is disputed. The company resisted calls to fact-check political misinformation for years [142], but according to anonymous Facebook employees, it actually did employ a "kill switch" for its algorithm from November 3-8, 2020 to prevent a US Presidential candidate from falsely declaring victory. This setting demoted the rankings of news sources Facebook deemed untrustworthy, and the so-called "nicer newsfeed" resulted in a decrease both in misinformation but also in user engagement/sessions. By the end of the month, the algorithm was essentially reset to its previous setting, because according to one employee, "the bottom line is that we couldn't hurt our bottom line. Mark still wanted people using Facebook as much as possible, as often as possible" [87].

When a tech company does not live up to its purported values, employees can serve as a meaningful check on the company's actions. Labor is well suited to be a countervailing force, because employees are relevant authorities with their technical expertise and knowledge of company goings-on. As issues arise, employee activism and collective power can be utilized to prevent technology companies from negatively impacting society.

In this project, I explore the use of tech worker collective action as a legitimate lever for doing AI Ethics: pushing companies to act in line with their stated values. First, I outline related work both in algorithmic accountability and social movements. Next, I profile three case studies of worker-led campaigns to push tech companies to avoid building harmful products. Using frameworks from political scientist Gene Sharp and labor organizer Jane McAlevey, I analyze an archive of hundreds of documented tech worker collective actions over the decades. I then discuss aspects that distinguish tech worker organizing from other forms of labor organizing. Finally, I use the Data Feminist principle to "make labor visible" to interrogate this overview for its own "project evidence."

5.2 Background and Related Work

5.2.1 Accountable Algorithms

Audits and case studies critically examine a system to determine if it is functioning the way it was intended and advertised. In foundational work for the algorithmic fairness community, Buolamwini and Gebru demonstrated bias in commercial facial recognition software towards women, towards people with dark skin, and towards the intersections of those groups [43]. Chouldechova et al. audit an algorithm-assisted child maltreatment hotline screening system and identify many of the challenges in implementing such an investigation in practice [60]. Yang et al. demonstrate how political censorship of Wikipedia can affect the pre-trained models used for general domain NLP algorithms [316]. Bender and Gebru et al. critically examine the environmental and financial costs first of large language models and offer some recommendations for curating and documenting datasets more carefully [22]. In an audit of a non-computational system, Cheong et al. examine the citation networks of many computer science-related fields and demonstrate that members are under-citing researchers from marginalized backgrounds (e.g., women) and recommend that the Association for Computing Machinery (ACM) has a duty of care to address this problem [58].

Excitingly, FAccT has increasingly been embracing theories of change beyond problem identification, computational methods, and philosophical discussions. Gebru et al. and Mitchell et al. introduce Datasheets for Datasets [98] and Model Cards for Model Reporting [185], respectively, for standardizing the transparency of algorithmic system development. Going one step further, Raji et al. develop a framework for algorithmic auditing to be applied throughout the internal organization development life-cycle and discuss the challenges of maintaining an independent and objective viewpoint during the execution of an audit [235]. Vincent et al. explore ways for users to influence tech companies through *data leverage*, where the users of a system "threaten[] to engage in or directly engag[e] in data-related actions that harm that organization's technologies or help its competitors' technologies" [292]. An interdisciplinary group from Computer Science departments, Sociology departments, the ACLU of Washington, and many other organizations built the Algorithmic Equity Toolkit, a set of reflective tools to increase public participation in technology advocacy for AI policy action [152].

In "Activism in the AI Community," Belfield observes the role that tech workers have played in shaping the societal and ethical implications of AI [21]. However, Belfield only engages with a handful of examples: Googlers resisting Project Maven, Googlers resisting Project Dragonfly, Googlers opposing workplace sexual harassment, and tech workers from many firms opposing corporate partnerships with Immigration and Customs Enforcement (ICE) and Customs and Border Protection (CBP). These examples, while high profile, represent a very narrow view of tech workers organizing for ethical computing practices. Additionally, that work largely considers industry-wide factors such as low union density and the widespread use by tech companies of non-disclosure agreements (NDAs). It does not yet form a convincing theory of why some campaigns succeed and others fail. In order to understand that further, we look to the literature on social movements and labor organizing.

5.2.2 Social Movements and Labor Organizing

Social change is the product of structural determinants (e.g., population change) and processes and mechanisms (e.g., political conflict and accommodation) [108]. Many philosophers, economists, historians, and political scientists have characterized different theories of change. Stephan and Chenoweth found that when resisting an oppressive government, nonviolent social movements are twice as likely to succeed than violent campaigns, and similarly that nonviolent movements are more likely to peacefully transition to a stable democracy. Through quantitative and qualitative analysis, they conclude this is because nonviolent methods allow for larger and more diverse movements, which engender increased resiliency, flexibility of tactics, and loyalty shifts from cross-pressured powerful actors [273].

Dr. Gene Sharp is one of the most influential theoreticians of nonviolent action; his methods were influential to pro-democracy campaigns in Serbia, Georgia, Kyrgyzstan, and Belarus. His work rejects the belief that people are fundamentally dependent upon the good will of their governments, and instead argues that governments are fundamentally dependent on "the people's good will, decisions and support" [259]. In his 1973 book 'The Politics of Nonviolent Action," he explores the theory behind nonviolent resistance; its success does not rely solely on persuading the opponent but rather often by persuading the other stakeholders on whom the opponent depends. He enumerates[1] 198 kinds of nonviolent actions (e.g., letters of opposition, singing, etc.) to demonstrate the power of a movement and pressure the opponent [259].

Sharp's enumeration of tactics is extensive though certainly not comprehensive; other resources also enumerate organizing tactics as well as describe how to perform an action in greater depth. For instance, the national rank-and-file union the United Electrical, Radio and Machine Workers of America (UE) host a public strike guide which gives high-level advice for how to plan for a successful worker's strike,[2] including by forming the right committees, setting up food distribution and travel for workers, ensuring utilities and rent/mortgage assistance, obtaining legal expertise, and more.

[1] https://www.aeinstein.org/nonviolentaction/198-methods-of-nonviolent-action

[2] https://www.ueunion.org/strikes

Unions have a rich history of “bargaining for the common good” [12], which is an approach of using contract fights to organize local stakeholders to fight for demands which would benefit people beyond the bargaining unit. For instance, after hosting community listening sessions, the 2018 LA teachers union strike included demands for green space for children and an immigrant defense fund for parents [101].

After decades of successful trade union organizing, Jane McAlevey got a PhD analyzing US labor movements in the 21st century [178]. Her work connects social movements with labor organizing and argues that the two do not have a clear distinction. She argues that democracy in the workplace is one of the most effective tools available to ordinary people for social progress, like the US saw in the labor movement in the 1930s-1940s and the Civil Rights Movement in the 1950s-1960s. Her analysis identifies the strategies, methods, and discipline behind successful and unsuccessful campaigns. We explore her work further in Section 5.5.

5.3 Tech Worker Campaigns: Three Case Studies

In 2017, Ossola pointed out that for industries like medicine, the government vets and tracks tools susceptible to abuse, in contrast with the tech industry, where that responsibility falls to individual companies [214]. However, after a series of scandals, it does not seem like companies are living up to their stated values of privacy [296], security [173], fairness [73], or safety [263]. In the past 5 years, tech workers have taken on a more active role than in previous years in discussing the social impact of their companies’ products.

Often times, meaningful channels for change do not already exist, and employees must organize and pressure their employer to take such actions. In this work, we refer to a goal-oriented, long-term effort as a “campaign,” which is composed of a series of individual “actions.”

The most comprehensive collection of such actions, to our knowledge, can be found at the Collective Action in Tech (CAIT) archive [282]. This project was created by former tech workers, union organizers, and a sociologist “to create a space for us to reflect on the tech worker movement's past, and invent its future." This archive is not guaranteed to be comprehensive and it largely consists of external vantages of how tech worker campaigns played out. Nonetheless, we notice some chronological trends as certain political issues increased in salience.

In this section, we highlight three successful tech worker campaigns: opposing a Muslim registry industry-wide, opposing a Department of Defense contract at Google, and opposing facial surveillance as a service.

5.3.1 Muslim Registry (2017)

While on the campaign trail in November 2015, then-candidate Donald Trump was asked if he would implement a database system tracking Muslims in the United States. He responded "I would certainly implement that. ... There should be a lot of systems, beyond databases. We should have a lot of systems" [119]. After he won the 2016 election, his transition team suggested the administration may pursue "extreme vetting" of some immigrants and bring back a Bush-era surveillance program (National Security Entry-Exit Registration System) which had been criticized for targeting immigrants from Muslim-majority countries (Of the 25 counties on the list, 24 were Muslim-majority, plus North Korea) [245] .

Many became increasingly worried that the Trump administration would follow through on its campaign promises to build a Muslim registry. In a (rare at the time) direct challenge to their employer, a group of over 50 IBM engineers authored a public letter calling for the firm to allow employees to "refuse participation in any U.S. contracts that violate constitutional and civil liberties" [25].

Tech workers launched NeverAgain.tech, which pledged to resist attempts to build databases to target individuals based on religion or national origin. 2,843 tech workers signed the pledge, including employees from Amazon, Apple, Google, and Microsoft.[3] Before the pledge, Twitter had been the only large tech firm that publicly opposed a Muslim registry, but after the Never Again campaign, there were also similar commitments from Facebook, Apple, Google, Twitter, IBM, Microsoft, Uber, Lyft Medium,

[3] `http://neveragain.tech`

and Salesforce [244]. After the tech community drew a clear, bright line about refusing to build a Muslim registry, the Trump administration did not pursue that specific policy.

5.3.2 Project Maven (2018)

In April 2017, the Department of Defense (DoD) established the Algorithmic Warfare Cross-Function Team to accelerate DoD's integration of big data and machine learning. As part of this effort, Google signed a contract with DoD for Project Maven, a $9 million project to build computer vision for drones, which was seen by many as a trial run for the much larger $10 billion JEDI contract [80].

When Google employees learned of Project Maven, many were concerned about whether Google was getting into "the business of war." Employees wrote a petition [258], signed by 4,000 Googlers, calling for the company to "cancel the Project Maven contract and publicly state Google and contractors will never build tech for war." After much discussion on Google's internal messaging boards and public pressure from media attention, the company executives hosted a discussion for Googlers to view in April 2018 between themselves and some of the petition authors. The town hall did not ease the concerns of employees, and frustrations began to mount.

In order to compete for secure government contracts, Google needed to implement "air gap" technology so that there would be physical separation between machines with government data and other machines. The influential group of software engineers tasked with building that tech for Google surprised their bosses by refusing to work on it [28]. They became known as the "Group of Nine" amongst their colleagues at Google, and their refusal increased pressure on the firm, which did not want to alienate or circumvent those influential engineers. However, without this tool, Google would be at a major competitive disadvantage in bidding for defense contracts against Amazon and Microsoft.

In June 2018, Google announced that it would be dropping the Maven project (i.e., declining to renew its contract the following year). A week later, Google announced the new Google AI Principles [229]. These principles include some abstract values, but also a few conceptual areas for which Google claims it won't pursue or deploy AI, such as "[w]eapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people."

5.3.3 Face Surveillance

Amid mass protests across the US in support of Black Lives Matter and criminal justice reform in 2020, many companies (Amazon, Microsoft, and IBM) suspended the sales of facial surveillance services [284]. But how did this happen? It took years of scholarship and activism [111] to get these companies to the point where that choice was the "safe" option, at least for the time, including:

- **October 2016**: Academic researchers (Garvie, Bedoya, and Frankle) published "The Perpetual Lineup" which warned that law enforcement agencies are using unregulated facial recognition technology to be able to surveil over 100 million Americans [96].
- **February 2018**: Academic researchers (Buolamwini and Gebru) published "Gender Shades," which found that computer vision models performed worse on dark skinned and female subjects [43].
- **May 2018**: The ACLU and a coalition of 48 civil rights organizations called on Amazon to stop allowing governments to use their Rekognition software in 2018 because the company's materials describe "person tracking" as an "easy and accurate" way to investigate and monitor people, such as undocumented immigrants or Black activists [47].
- **June 2018**: Citing the ACLU report, 500 Amazon employees signed an open letter[4] calling Amazon to "stop selling facial recognition service to law enforcement" and to "stop providing infrastructure to Palantir and any other Amazon partners who enable ICE."
- **July 2019**: A group of Amazon employees sent an email to internal employee mailing lists, demanding that Palantir be removed from Amazon's cloud for violating its terms of service and for Amazon to take a stand against ICE by

[4] `https://www.scribd.com/document/382334740/Dear-Jeff`

making a statement [54].

- **June 2020**: After tens of millions of protesters took to the streets over the murder of George Floyd by police, IBM announced it would discontinue selling facial recognition software. The following day, Amazon announced a one-year moratorium on police use of Rekognition.
- **May 2021**: Amazon announced that it would indefinitely prohibit police departments from using Rekognition.

Unlike individual firm campaigns that take place at one single company, this industry-wide effort was able to stigmatize the unregulated use of this new technology enough that it changed the market. For years, many scholars and activists had worked to slow the development of unaccountable facial surveillance technology, but mostly did not "move the needle" on company priorities. However, when the 2020 Black Lives Matter protests demonstrated energy for change, many corporations reached for the solutions that were fleshed out and based upon research.

Of course, without legislation, this is still a live issue wherein vendors could decide to reverse course and begin production again if they no longer fear the potential backlash.

5.4 Tech Worker Actions: A Wider Analysis

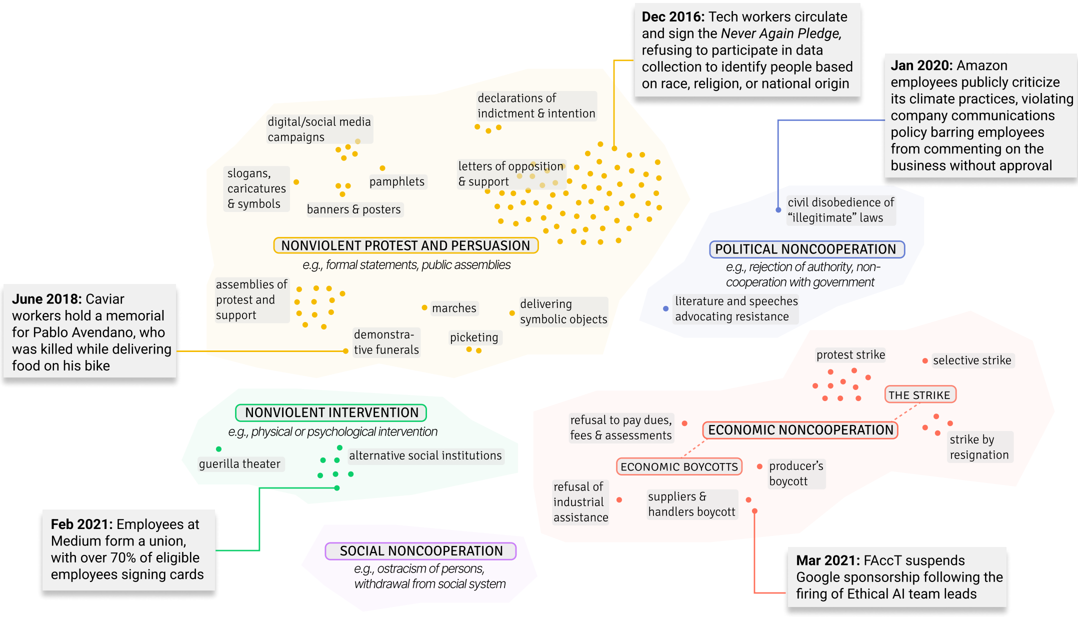
In this section, we provide a more systematic analysis of tech worker collective actions. We categorize a set of 150 actions from the CAIT archive [282] into Sharp's framework of nonviolent actions. Our goal is to understand the broader space of tech worker collective actions in recent years, both to examine actions that have been widely utilized as well as demonstrate the much broader space of possible actions to explore.

Sharp's framework categorizes the methods of nonviolent action into a few broad categories, including nonviolent protest and persuasion, social noncooperation, economic boycotts, the strike, political noncooperation, and nonviolent intervention. Table 5.1 contains a subset of particularly relevant nonviolent methods that he enumerates.

Table 5.1: Subset of nonviolent methods enumerated by Sharp.

Overarching Methods	*Action Groups*	*Example Actions*
Nonviolent Protest and Persuasian	*Formal statements*	Letters of opposition, public speeches
	Honoring the dead	Mock funerals, political mourning
	Public Assemblies	Teach-ins, assemblies of protest
Social Noncooperation	*Withdrawal from the social system*	Stay-at-home, collective disappearance
Economic Noncooperation: *Economic Boycotts*	*Actions by consumers*	Consumers' boycott, non-consumption of boycotted goods
	Action by middlemen	Suppliers' and handlers' boycott
Economic Noncooperation: *The Strike*	*Strikes by Special Groups*	Craft strike, professional strike
	Restricted Strikes	Slowdown strikes, working-to-rule strikes
Political Noncooperation	*Citizens' noncooperation with government*	Refusal of assistance to enforcement agents, removal of own signs and placemarks
Nonviolent Intervention	*Physical intervention*	Sit-in, nonviolent occupation
	Social intervention	Overloading of facilities, alternative communication system)

Figure 5-1: Depiction of the collective actions tagged as "ethics" in the CAIT archive. Each action was categorized into Sharp's framework. The axes of this figure do not signify quantitative meaning.


digital/social media campaigns
declarations of indictment & intention
slogans, caricatures & symbols
pamphlets
letters of opposition & support
banners & posters
NONVIOLENT PROTEST AND PERSUASION
e.g., formal statements, public assemblies
assemblies of protest and support
marches
delivering symbolic objects
demonstrative funerals
picketing
civil disobedience of "illegitimate" laws
POLITICAL NONCOOPERATION
e.g., rejection of authority, non-cooperation with government
literature and speeches advocating resistance
protest strike
selective strike
THE STRIKE
refusal to pay dues, fees & assessments
ECONOMIC NONCOOPERATION
strike by resignation
ECONOMIC BOYCOTTS
producer's boycott
refusal of industrial assistance
suppliers & handlers boycott
NONVIOLENT INTERVENTION
e.g., physical or psychological intervention
guerilla theater
alternative social institutions
SOCIAL NONCOOPERATION
e.g., ostracism of persons, withdrawal from social system

Dec 2016: Tech workers circulate and sign the *Never Again Pledge*, refusing to participate in data collection to identify people based on race, religion, or national origin

Jan 2020: Amazon employees publicly criticize its climate practices, violating company communications policy barring employees from commenting on the business without approval

June 2018: Caviar workers hold a memorial for Pablo Avendano, who was killed while delivering food on his bike

Feb 2021: Employees at Medium form a union, with over 70% of eligible employees signing cards

Mar 2021: FAccT suspends Google sponsorship following the firing of Ethical AI team leads

To categorize collective actions according to Sharp's framework, we first filtered the CAIT archive by those actions tagged with "ethics", in order arrive at a set of 139 entries more relevant to our focus.[5] Each action was tagged by me or one of my co-authors, and cases where there was uncertainty or disagreement were solved through joint discussion and further research into the particular event. If a particular entry in the archive seemed to describe multiple actions (e.g., an event involving both a letter of opposition and a protest strike), we considered that as two separate actions for the analysis. This resulted in 150 actions in the final coded archive. Fig. 5-1 depicts the summarization of each action and its categorization.

Overall, we found that Sharp's framework was broad and detailed enough to categorize the range of tech worker collective actions described in the CAIT archive. However, there were also entries in the archive for which there was not an existing category in the Sharp hierarchy. In many cases, these indicated innovative avenues

[5] The full table with each action and our codes can be found at `https://docs.google.com/spreadsheets/d/1QDFJiNvwYL-MFdfS5ZobDB9DBFITHi-Uu6J7dkUfeio/edit?usp=sharing`

for collective action that are opened due to modern technology and/or the nature of the tech industry (e.g., social media campaigns, pressure from company shareholders, virtual walk-outs via “closing laptops”). There were also a few types of actions in the archive not covered by the framework—in particular, because Sharp focuses primarily on labor power, actions such as lawsuits that utilize other forms of power (i.e., legal power) are not covered.

Examining the distribution of actions, we found that letters of support/opposition and group petitions make up the majority (n=73). Actions such as assemblies of protest or support (n=13), protest strikes (n=11), and alternative social institutions (e.g., unions, n=9) are much less common, but have still been moderately explored in different contexts. Most of Sharp’s other 198 actions types have not been explored, or have just one or a handful of instances (e.g., suppliers and handlers boycott, guerilla theater, civil disobedience of “illegitimate” laws).

While not all of the actions in the CAIT archive were effective, there are many examples of successful demonstrations of collective power. Here, we highlight some specific instances of strong actions (many of which were situated in broader movements or campaigns), demonstrating the type of action(s) utilized and how they fall within Sharp’s framework (numbers in parenthesis indicate the number of the corresponding action in Sharp’s full framework linked to in Section 5.2.2).

Never Again Pledge: In December 2016, a group of tech workers circulated an online pledge refusing to participate in developing technology or collecting data that could aid in identifying people by race, religion, or national origin. The pledge was specifically created in response to the Trump presidential campaign’s comments around creating a “Muslim registry.” 2,843 tech workers signed the pledge, which created a significant amount of media coverage, public attention, and spurred statements of refusal from a range of tech companies. In Sharp’s framework, this action fall under *letters of opposition or support* (#2) and *group or mass petitions* (#6). It also utilizes *slogans, caricatures, and symbols* (#7) — i.e., the strong rhetoric of “Never Again.”

Industry Refusal to Build Muslim Registry: In part spurred by the Never

Again Pledge, there was an effective industry-wide effort to resist building surveillance tech against religious minorities. This action is consistent with Sharp's *refusal of industrial assistance* (#84) as well as *boycott of government depts., agencies, and other bodies* (#126).

Caviar Memorial: In June 2018, after a gig worker died during a delivery for Caviar, fellow gig workers organized a memorial and raised money for the funeral [223]. They demanded Caviar pay for the funeral expenses, classify riders as employees (not independent contractors), give a starting salary of \$20/hour, and respect workers' rights to organize a union. Following this, in July 2018, Caviar began offering accident insurance to all driver actively picking up or delivering an order. This action was the only example we encountered in the archive that utilized *demonstrative funerals* (#45).

FAccT 2021 Dropping Google as Sponsor: Between December 2020 and Febrary 2021, Dr. Timnit Gebru and Dr. Margaret Mitchell—the co-leads of Google's Ethical AI team—were fired from Google. In response, the FAccT research community suspended Google's sponsorship of the FAccT 2021 conference [134]. There was not an immediate demand associated with the action, but a reasonable interpretation is that it was taken to act as a deterrence for similar behavior from companies in future situations. This action is an example of a *suppliers' and handlers' boycott* (#80).

Amazon Worker-backed Shareholder Resolution: In April 2019, Amazon employees publicly supported a shareholder resolution requesting that Amazon's Board of Directors "prepare a public report as soon as practicable describing how Amazon is planning for disruptions posed by climate change, and how Amazon is reducing its company-wide dependence on fossil fuels" [61]. Although the resolution was voted down by shareholders (70% opposed, 30% supported) Amazon Employees for Climate Justice continued with collective actions, leading to Amazon creating The Climate Pledge to meet the Paris Agreement 10 years early. This shareholder resolution tactic does not appeal directly to labor power, but rather through public pressure targeting the shareholders. This resolution did not immediately require ratification in order to be a successful demonstration; by pressuring the shareholders through

Delivering symbolic objects (#21), workers demonstrated a diversity of tactics for pressuring Amazon.

Exercising Legal Workplace Protections: In addition to labor power, workers sometimes also have legal power on their side. In April 2020, Amazon fired two of its tech workers after they publicly criticized the company's warehouse workplace conditions amid COVID-19 [105]. The employees alleged their terminations were retaliation for their advocacy around both coronavirus working conditions and climate advocacy at Amazon. Although employers often have more resources — and therefore are better able to draw out long legal battles — when employees do exercise their legal rights, it can serve as disincentive for companies considering pushing the line of legal gray areas. Increasing the chances that inappropriate behavior actually would lead to legal battles arms employees with a credible threat and serves as a check on employer power. The employees settled their case with Amazon, receiving an undisclosed amount of money [106]. Because this action utilizes legal power (i.e., a lawsuit), it is not described in Sharp's framework.

5.5 Building Tech Worker Power

A successful action (and especially a successful campaign) does not just happen on its own. There is almost always a lot of invisible labor that goes into building the social infrastructure to coordinate large collective actions. In this section, we examine the methods of prominent labor organizer Dr. Jane McAlevey and discuss how those concepts could be applied to tech.

McAlevey analyzes many union campaigns, and her conclusion for how to win a strong contract is not complicated: 1) map out an honest theory of power, 2) create a credible plan to win, and 3) execute that plan with strong methods and discipline [178]. She argues that most modern social movements fail to develop a theory of power, and as a result, unwittingly set themselves up to fail. Critically, she emphasizes that many people involved in social movements conflate two importantly distinct concepts: "organizing" and "mobilizing." A lot of people *think* they are doing organizing when in reality they are only talking to people who already agree with them. Organizing, on the other hand, involves bringing new people into the campaign and growing the base of collective power from which one can mobilize later. Going back to the theory of power, McAlevey concludes that if a campaign can "creat[e] a crisis for the employer" then they will win and if they cannot do that then they will lose. She analyzes many union campaigns (some successful and some unsuccessful) and demonstrates that the only way to win hard fights is to do genuine organizing in order to build super-majority support of workers. Only with that level of support can workers have a credible threat (e.g., by striking) in order win a strong outcome.

The first two steps of McAlevey's process for a successful campaign are to map power and set a corresponding credible plan to win. Depending on how difficult the goal is, different tactics will need to be employed: a minority of vocal workers could win the "Justice for Janitors" campaign because concessions were low-cost to the employer, but in order for hospital nurses to win costly nurse-to-patient ratios, a nursing strike may be required [178]. As she summarizes: "High concession costs require high power."

The third step of the process is to execute that plan using effective methods and discipline. But what are those methods? In this section, we distill her organizing concepts and discuss how they can be applied to the tech sector.

5.5.1 McAlevey: Organizing for Power

McAlevey runs trainings for how workers can build power and win the material benefits they'd like to see.[6] She contextualizes the points above, working backwards from understanding what it would take to build a movement strong enough to win. Recognizing that workers' power doesn't come from money or status, but instead from large numbers and taking collective actions together, she discusses the methods and disciplines of successful labor organizing campaigns. The fundamental unit of work is the one-on-one "organizing conversation" between two coworkers, where the organizer

[6] https://podcasts.apple.com/us/podcast/a-master-class-in-organizing/id1081584611?i=1000468514310

identifies their colleague's issues and connects them to solutions rooted in collective action.

Throughout her trainings and scholarship, she identifies some critical concepts that successful organizers use. For each one, we demonstrate corresponding examples in the tech sector. Sometimes, such as when describing a campaign's strategy, we are left to speculate and make reasonable guesses about the intention of the organizers behind the action/campaign.

Issue Identification: Organizers must identify what issues are important to their coworkers during a one-on-one by asking open ended questions. An effective way to identify actionable issues they wish would improve is by asking "If you could change 3 things at work tomorrow, what would they be?" By identifying their priorities, organizers can then discuss how those issues connect to collective solutions by coworkers with similar concerns.

Many high-profile campaigns in tech were in reaction to a big political event, such as the Never Again Pledge in response to the Trump campaign suggesting a Muslim registry [245]. In other campaigns, tech workers came to understand that they had the power and responsibility to solve a problem even if the issue wasn't front-and-center politically. For both Project Maven and Face Surveillance, many employees who ordinarily did not want to "rock the boat" felt that they had to be part of the solution, because if not them, then who?

Raising Expectations: People will not fight for more unless they believe that they deserve more and that they could actually have more. One of the most effective ways to convince people that things can be better than they are now is by showing successes elsewhere.

For instance, in March 2021, the Glitch union signed the first collective bargaining agreement for software engineers [154]. The contract did not set wage floors or salary rules but instead focused on "protecting basic labor rights, challenging discriminatory pay and hiring practices, and even pushing companies to be held accountable for the products they build." Less than a month later, when the workers at Mobilize announced the formation of their own union, they pointed to the Glitch workers as an example they took inspiration from: “Like Glitch, I think that we can serve as an example for other employers to see ... that we can work together to figure out what workers want” [153].

Credible Plan to Win: Organizers must do a power analysis to understand the concession costs associated with their goals, and then plan how to generate enough power to achieve that success. As McAlevey observes in her dissertation “An incorrect power analysis might lead people who want to end capitalism to think that small numbers of demonstrators occupying public spaces like parks and squares and tweeting about it will generate enough power to collapse Wall Street” [178].

Creating this credible plan will involve mapping out a set of plausible steps which can ladder up to a successful campaign. For instance, looking at the Never Again pledge, organizers were able to pressure individual firms one at a time. Each time another firm made a public statement, it served as a domino, making it easier for the next firm to make an announcement too. Eventually, the campaign was able to build a consensus around industry-wide opposition to the proposed Muslim registry.

Structure-based Organizing: Within any workplace, there are already existing structures and social networks — perhaps by floor, by department, by communities/clubs, etc. These structures have existing social dynamics and relationships of trust which one should organize within, as opposed to trying to build an entirely new structure from scratch.

One (but by no means the only) opportunity for identifying coworkers with mutual interests is through Employee Resource Groups (ERG). The modern ERG emerged as part of the civil rights movement when Xerox workers created the National Black Employee Caucus to “push back against racist business practices and systems” [191]. ERGs are quite common in tech companies; Google has 16 ERGs for nearly 25% of the workforce (35,000 of 140,000 as of 2021) [69]. In 2019, workers on Google’s LGBT ERG (Gayglers) organized a petition to pressure Google to change its policy on YouTube’s moderation decisions affecting the LGBTQ+ community [81].

Organic Leader Identification: Many unsuccessful campaigns are lost because the wrong leaders were selected, causing them to be out of step with the broader membership and make decisions which fail to attract buy-in. Within any given structure, there are the most trusted members of that group. Naive questions like "who is your leader?" or "who do you respect most?" often lead to incorrect leader identification because words like "leader" are imprecise and people's plain-use understanding of the word might not lead to thinking of the right leaders for the campaign. Just because someone shows up to meetings or gives a good speech, that does not mean a majority of their coworkers trust them. Instead, organizers can identify the organic leader by asking a majority of the members of the group questions like "If your manager asks you to do something, and you're not sure how to do it, who do you ask for help from?"

Structure Tests: In order to measure the health of the campaign (including whether one has identified the organic leaders in the relevant units), organizers should run a series of structure tests to see if their organizing network is as strong as they think it is. First, organizers decide on an action they'd like everyone to take, and then second, they communicate that through their action network. By gauging participation from each subdivision of their organization, they can measure their capacity for mobilizing. The goal is to realize where improvement is necessary *before* flexing collective power publicly. Is there a given floor, department, or team where participation in the structure test is much lower than average?

Examples of structure tests include majority petitions, photo posters, sticker days, wearing t-shirts with union emblems, and rallies. For example, the Alphabet Workers Union has a zoom background which workers can use [288]. Additionally, the United Auto Workers (UAW) and Communication Workers of America (CWA) encourage locals to participate in events like "Red Shirt Wednesdays," where members pledge to all wear red on a given day.[7]

Framing the Hard Choice: During an organizer's one-on-one organizing conversation, their colleague might agree that a problem exists, but perhaps they are hesitant to take a stand about it. The important thing is for them to come to the conclusion themself about who has the power to change the situation and whether that person will ever do that without being pressured to do it by their employees.

[7] https://uaw.org/wp-content/uploads/2016/06/Red-Shirt-Wednesday.pdf

McAlevey recommends leveling with the coworker and demonstrating that the organizer shares their concern but then asking how else the issue will be resolved unless employees band together and stand up for what is right.

For instance, in May 2020, then-President Trump incited violence against protesters with his "when the looting starts, the shooting starts" dog whistle on social media, which echoed statements of Walter E. Headley and George Wallace [44]. Twitter and Facebook both chose to keep the post up; Facebook in particular took no action at all despite it violating Facebook's previously stated community guidelines, leading to employees' survey-reported confidence in Facebook leadership to plummet from 75% to 47% and pride in the company from 73% to 48% in a matter of weeks [216]. After the January 6 insurrection happened, Facebook and Twitter employees no longer trusted management to address the problem without being pushed [138]. A potentially effective framing could be: "Every day nothing happens is another day of hurting our users. If we don't band together to do something, how else is this ever going to change?"

Inoculation: Another way campaigns can fail is if the organizers do not adequately prepare to withstand management's tactics to undermine the movement. This is not hypothetical; employers hire outside consultants[8] that specialize in sowing confusion and fear to get employees to second guess taking collective action. During the end of the first one-on-one conversation, McAlevey recommends "giv[ing] the worker a little bit of the 'poison' they will hear from management" in order to reduce the anxiety for when it does happen. This can be accomplished by asking something like "do you think your boss is going to like it if their employee signed an open letter calling for change? Why?"

Although this might sound like overkill, tech companies have been ramping up professional efforts to undermine tech worker organizing. In 2019, Google hired IRI Consultants, a top union busting firm in the United States whose website advertises their success in avoiding labor organizing and in conducting union vulnerability assessments [252]. According to reporting, Amazon uses a "heat mapping" tool to

[8] https://www.youtube.com/watch?v=Gk8dUXRpoy8

identify Whole Foods stores at risk of unionizing based on factors including the number of complaints filed to the NLRB, the poverty rate for the store's zip code, the racial and ethnic diversity of a store, the average employee compensation, and how employees felt about their workplace [228].

Stakeholder Organizing: When the organizer is mapping power as part of the credible plan, they will encounter additional stakeholders on whom the firm depends (e.g., customers, vendors, positive media coverage, school internship pipelines, etc). Successful campaigns are often able to build connections with other stakeholders to coordinate putting pressure on management from multiple directions.

We can see an example of this from a pressure campaign on Microsoft-owned GitHub. In October 2019, hundreds of GitHub employees signed an open letter calling on the company to cancel its contract with ICE [99]. In December 2019, over 700 developers that use GitHub also signed an open letter supporting the workers' calls for GitHub to cancel its contract with ICE [287]. However, GitHub currently boasts tens of millions of developers using its platform, and the amount of power that the campaign amassed was not enough to win the concession cost of dropping the contract with ICE.

Credible Threat: The credible threat (as demonstrated by a successful structure test or previous action) is the leverage for the campaign to bargain with the employer. The most clever plan or rhetoric in the world is not a substitute for whether the employer looks across the bargaining table and sees a super-majority of the workers saying "We don't want to go on strike, but we are prepared to if our needs are not met."

In November 2018, more than 20,000 Google employees (over 25% of the workforce) participated in a worldwide walkout to protest how Google handled cases of sexual harassment [297]. They demanded transparency, the presence of an employee representative, and the public filings of each sexual assault case. As a result, the company published an internal report of sexual assault cases, and in February 2019, ended the practice of forced arbitration.

5.6 Discussion

There is a lot we can learn from both previous instances of tech worker collective action and theories for organizing and social movements.

5.6.1 Theory of Power

As McAlevey observes, campaigns are won and lost based on whether the leaders had a good strategy (aka "credible plan to win"). It is important to understand who has the power to make the desired change and how much rank-and-file power would need to be built to push them [178]. The instances from the CAIT archive demonstrate examples of how tech workers have utilized different forms of power: public pressure (e.g., public letters), legal power (e.g., suing companies for violating labor laws), shareholder power (e.g., workers backing a shareholder resolution), labor power (e.g., walkouts), and more.

To that end, we observed from Section 5.4 that there was a very large number of open letters, internal letters, and petitions. Of course, these are an important part of demonstrating collective power and rallying external stakeholders to one's cause, but their power is less strong when they are not accompanied with a (either explicit or implicit) credible threat. Walkouts and protest strikes (e.g., Googler's 20,000-person walkout about sexual harassment) are a closer demonstration of commitment from the workers to cause sufficient disruption to win demands with large concession costs. There are not yet examples of a majority strike in the archive.

There were instances of some early successful actions, especially in 2017-2018, where a vocal minority (1-3%) of the workers created public pressure and were able to win their goals. However, in the years since, companies have also learned from these examples and adapted their responses so that they are more willing to take a cycle of bad press and try to wait the campaign out. This is in line with a similar finding in social science theory, where Dr. Leah Cardamore Stokes observes that new tactics or policies often start with a "Fog of Enactment" where powerful incumbents do not initially understand the impact of a new policy and they miscalculate how to respond [275]. Eventually, however, the incumbent learns how to more accurately assess the policy and is able to counter it more effectively in later efforts.

Ultimately, companies hire workers because labor keeps the company running. Organizing a majority strike is still a gold standard of leverage. Jerry Brown, the retired President of 1199NE, said "[t]he strike muscle is like any other muscle, you have to keep it in good shape or it will atrophy." Under Brown, his union of Connecticut nursing home workers went on strike over 100 times and won a large number of their bargaining goals [178].

5.6.2 Expectations and Timeline

In order to build the requisite amount of people power to win hard campaigns for more ethical products, tech workers will need to organize.

The most traditional model of labor organizing involves getting a large group of workers to participate in collective actions, where the ultimate leverage comes from the threat of a strike. In order to create a such a large campaign, organizers must raise the expectations of people who currently believe either the status quo is good enough or even if it is not that it won't change. Identifying the bright spots of other successful campaigns can be an effective way to show what else is possible, and help tech workers understand the power they might not have realized that they have [117].

One challenge with raising expectations is that workers run the risk of getting their hopes up for what is possible only to run into disillusionment if they are not able to achieve everything they want. By looking to previous efforts, we see that successful campaigns require sustained action: tech workers continued pushing against facial surveillance tools for years in the forms of academic scholarship, internal letters, and open letters. The years of effort have (thus far) paid off because when external forces (i.e., millions of Americans marching for racial justice in 2020) put additional pressure on Amazon, they reached for the solutions that organizers had spent years engineering.

5.6.3 Tech-specific Considerations

Although we employ Sharp's framework for categorizing the many forms of collective actions, this ontology does not perfectly reflect the state of tech worker organizing. We hope that the large action space of Sharp's methods can serve a generative purpose to help identify action types which have not yet been attempted in tech but could prove useful. Additionally, there are some types of actions which are unique to tech and thus not included in Sharp's general framework. We explore some of those considerations here.

One recurring tactic not captured in the Sharp framework is expert assessment of feasibility. In 1986, dozens of technical experts, including Herbert Simon (recipient of both the Nobel Prize and Turing Award) and John Backus (the inventor of FORTRAN), came out against President Reagan's "Star Wars" defense program, on the grounds that it was technically infeasible to build and test such a complex system for something as high stakes as a bug-caused nuclear strike [36]. In the 1990s, the Clinton administration pushed for Clipper Chip technology to allow for law enforcement to access encrypted data, but tech experts argued[9] that the technology was too technically flawed and insecure [1]. Most recently, for the past 5+ years, machine learning experts have been cautioning against sweeping deployment of ML — including facial surveillance tools — because the algorithms often exhibit racial biases [10]. Sharp's framework does not presuppose its practitioners are domain experts, so it does not explore the ways expertise can enable additional forms of public pressure on decision-makers.

To this end, McAlevey also recognizes that power is not always evenly distributed across all workers. While typically a feasible theory of power might require a supermajority (e.g., to pull off a strong strike), she also provides examples of successful campaigns that utilize critical workers [178]. In her dissertation, she analyzes a union drive at a Smithfield Foods pork factory where Livestock was a "key department" because workers could stop letting hogs off trucks, which both stopped the factory line and caused a massive traffic blockcade on the major interstate highway. A high-

[9] `http://cpsr.org/prevsite/program/clipper/cpsr-electronic-petition.html`

impact action didn't need the entire factory organized in order to work, it would just need to start with Livestock [178]. We expect to see many similar situations in the tech sector, where a small number of critical, specialized workers have an outsized effect on systems that are built. The closest example of this from the CAIT archive is the "Group of Nine" influential cloud engineers from the Google Project Maven campaign that refused to build the air gap technology. Although this is similar to the Livestock example, it does not fully capture the concept; reporting suggests "[the air gap] feature is not technically very difficult, so Google could easily find other engineers to do the work" [28]. Nonetheless, this serves as a potential blueprint for how an influential or specialized team can recognize and leverage their power in the tech industry. The most powerful groups will likely feel a sense of duty and reluctance to use that power carelessly, which can serve as a check on over-use. Just as "high concession costs require high power", conversely for low-power campaigns, there is no need to kill an ant with a sledgehammer.

Workers at a company usually would rather not "create a crisis" without a good reason, but how could they do that even if they wanted to? Different organizations have different power structures that determine which stakeholder's support is critical to the mission. For instance, gig workers have a traditional labor model wherein they could stop the service if they stopped working. On the other hand, there is not an immediate, acute harm to the organization if software engineers aren't patching bugs or building new features to compete with competitors. That harm becomes a long-term one which is harder to measure. Reporter Peter Kafka observes, "these companies live and die on their ability to recruit and retain top talent. That's a large part of what drives them to make these decisions" [138]. Companies are competing against each other for hiring top talent, and the fear of losing out contributed to the success of some early open letters. However, as that tactic has been used over the last 5 years, organizations have learned that if the only action will be an open letter, then they can wait the concern out without much cost. To effect change, workers will need to correspondingly organize additional kinds of collective actions. This suggests that one possible, relatively unexplored point of leverage could be exploring ways to interact with the company's recruiting, such as through unauthorized climate surveys or accountability scorecards that show responsiveness to employee requests.

5.7 Project "Evidence"

In this work, I move beyond individual computational case studies in order to study how tech workers have been able to practice AI Ethics beyond research and publications. Using Sharp's and McAlevey's frameworks, we explore the landscape of tech worker campaigns and actions to push companies to act more ethically. We can reflect on this effort by analyzing it against the Data Feminism principle to **make labor visible**. This concept applies in more ways than one, referring both to the work of doing this research project and also increasing the salience of labor organizing efforts. In the subsections below, we discuss how each of these relate to the broader impacts of ethical computing.

5.7.1 Reconstructing the Work of This Project

We begin by considering the labor that went into this research. First, other researchers created the archive documenting hundreds of tech worker collective actions. Then, we analyzed that archive using frameworks from political scientists and labor organizers.

The CAIT Archive

The Collective Action in Tech (CAIT) project began as a project to document those tech workers pushing their companies "in the right direction' using collective action [282]. It is a volunteer-run organization maintained by:

- Ben Tarnoff, a tech worker, writer, and cofounder of Logic Magazine
- Clarissa Redwine, former Kickstarter union organizer
- JS Tan, a former tech worker and writer
- Kristen Sheets, a tech worker and writer
- Nataliya Nedzhvetskaya, a sociologist who researches tech and labor and is supported by funding from the Jain Family Institute

The archive was constructed using NexisUni news archives, where the creators searched for articles about the computing and IT industry where employment terms (employee, worker, contract, labor) occured within 25 tokens of collective action terms (protest, petition, strike, open letter, walk out, union, boycott, letter, lawsuit, discuss, negotiat). For now, the effort primarily uses English-speaking news publications, though one of the co-founders writes about tech, labor, and China. As of 2021, approximately 5% of entries have been added through crowdsourcing [199].

Our Research

In this section, I recapitulate how this work was done. Although such work is usually left out of the public eye, there is value to this analysis both by making the goals of the project more explicit for reflection and also by allowing other researchers to reconstruct a similar effort for themselves if they want to build on the project. We want to use extracted "evidence" from projects to decide what projects were effective, and also *how* we could do conduct/promote similar efforts.

This project came together from the convergence of a few parallel efforts:

- Bianca Lepe and I were organizers for a 2020 collective action campaign for anti-discrimination;
- the December 2020 ousting of Dr. Timnit Gebru from Google sent shockwaves through the algorithmic fairness community; and
- I discovered the CAIT archive in Spring 2021.

As these efforts intersected in Spring 2021, Bianca and I began discussing a research project related to organizing.

Initially, the project could have gone in multiple directions, depending on whether the focus would be emphasizing tech worker organizing in general, diversity, equity, and inclusion (DEI) organizing, the relationship between diversity and ethical computing, and more. In July, we met with Dr. Catherine D'Ignazio for guidance on scoping the project into one that could be studied. We began to converge on the idea of using existing frameworks (e.g., McAlevey) to study the CAIT archive. The student team (Bianca, myself, and also eventually Harini) met weekly from July 2021 to January 2022, checking in with the Catherine at least once a month.

In late July, the team created a spreadsheet of the CAIT entries to better understand the lay of the land, coding each article for action, number of participants, demands, outcome, and quantified estimate of success where possible. We did not initially have a framework for categorizing this into a standardized form.

We then made a spreadsheet of all 2018 FAccT papers from 2018–2021. We read the abstracts of all 82 FAccT 2021 papers and selected other FAccT 2018–2020 papers to read based on relevant titles, authors, and citation counts. This landscape analysis helped us more confidently identify some of the conference's strengths (e.g., high-profile standards and recommendations such as Datasheets for Datasets [98] and the SMACTR audit framework [235]) and some areas for growth (e.g., a lot of the work relies on voluntary buy-in, which might not always be realistic). Data Leverage [292] from the 2021 conference was the closest to what we had in mind; it proposed a similar argument for corporate accountability but discussed what users could try to do, rather than workers.

The first framework we used to analyze the CAIT entries was the McAlevey tools for organizing. We read her dissertation, books [178], "organizing for power" training, and a 2020 interview.[10]

We initially were unsure how to visualize the collective actions. We sketched a 1.0 figure to iterate upon based on feedback. We were unsure what, specifically to plot and what the axes should correspond to (e.g., number of participating employees, intensity of action, degree of success, etc). The most common advice we received was to change the unit of analysis from *types* of action (e.g., "petition" or "email campaign") to specific actions (e.g., "the 2016 Never Again pledge").

In the Fall, we met with Harini to discuss effective ways to visualize the analysis and the actions. She brought theory and expertise to our previously ad hoc effort, and excitedly she joined the team. Her visualizations were much closer to what would eventually become Figure 5-1 after a few more rounds of feedback. One initial insight was to avoid the axes corresponding to specific measurements (e.g., degree of success)

[10] https://podcasts.apple.com/us/podcast/a-master-class-in-organizing/id1081584611?i=1000468514310

because the archive entries were news article that were far too imprecise and devoid of context to try categorizing and ordering all of the actions. Instead, the new figure was inspired by Crystal Lee et al.'s "Viral Visualizations" Figure 1 [157].

In November, a colleague in Political Science recommended a second framework for understanding the collective actions. Where the McAlevey tools help demonstrate how workers built power for the actions, Gene Sharp's "The Politics of Nonviolent Action" [259] categorized 198 types of actions. This framework became central to the revised analysis; we re-coded the archive entries into Sharp's categories and clustered the visualization's entries based on those labels.

The final pieces of the project came together through iterative writing and discussion sections during December 2021 and January 2022.

5.7.2 Increasing Awareness of Current Labor Organizing

The seventh principle of Data Feminism is to "make labor visible." Interpreting that guidance literally, this chapter aims to make *organized* labor more visible. But what does it mean for something to be visible? Is it a binary visible/invisible or a spectrum? And visible to whom? In this reflection, we interrogate these concepts further.

Visible to Whom?

To understand whether employee-driven ethical computing is already visible to the algorithmic fairness community, we explore how often these topics are discussed. Table 5.2 shows the 14 (out of 208 total) FAccT papers whose abstracts mention at least one employment or collective action term identified by the creators of the CAIT archive [282]. The two most cited works in that list (Chouldechova et al. [60] and Passi and Barocas [225]) both explore the role that workers have in shaping how computational systems should be designed in order to be useful and ethical. However, many of these papers have not received much traction in the community.

This collection demonstrates both that such a topic is within the domain of FAccT and that there is room to grow. Such a state represents an opportunity to increase

Table 5.2: FAccT papers (2018-2021) whose abstracts use the employment and collective terms developed by the creators of the Collective Action in Tech archive.

Type of Term	*Reference*	*Term*	*Citations*	*How It Was Used*
Employment	*Chouldechova et al. 2018*	worker	176	Considering the perspective/workflow of the workers at child maltreatment hotline that uses algorithm-assisted decision making
	Babaei et al. 2019	worker	24	Uses "Mechanical Turk workers" for annotations.
	Harrison et al. 2020	worker	38	Measured the fairness preferences of Mechanical Turkers when presented with 2 imperfect models.
	Terzis et al. 2020	worker	8	Discusses the philosophy of a multi-stakeholder (including CEOs and workers) "AI Ethics" round table, which it characterizes as "a futile battle doomed to dangerous abstraction."
	Jacovi et al. 2021	contract	49	N/A (referring to "contractual trust" between AI and user, not employer-employee)
	Miceli et al. 2021	worker	16	Interviewed 30 data workers in industry about data documentation and reflexivity.
	Celis et al. 2021	employee	2	Studying the effect of policy (the Rooney Rule) on hiring employees.
	Lussier et al. 2021	labor	0	A historical computerization case study how IP considerations, labor, technology, and expertise shaped the deployment of a program.
Collective Action	*Passi et al. 2019*	negotiat	80	Through six months of ethnographic fieldwork with a corporate data science team, examines how stakeholders (including workers) negotiate problem formulation and its ethical implications.
	Young et al. 2019	strike	32	N/A (says something "strikes a balance")
	Marcinkowski et al. 2020	protest	45	Found that data subjects (prospective students) were less likely to protest the results of an algorithmic decision-making than human decision-making from a study in German universities.
	Kaminski et al. 2021	union	26	N/A (refering to the European Union)
	Shen et al. 2021	negotiat	11	Value Cards help communicate how different models and deployment contexts have trade-offs to be negotiated amongst stakeholders.
	Mulder et al. 2021	strike	5	Studied framing effects of news media, gave "farmer's strike" as an example topic.

the salience of the role that data workers can play in defining the ethical trade-offs of a computational system. In the following section, we discuss visibility beyond a yes/no binary formulation.

Rethinking Visibility Beyond a Binary

Because this work's primary contribution is analyzing tech worker campaigns reported in public articles, technically speaking the work is already "visible" to anyone who looks for it. However, there is too much information on the internet for anyone to be able to follow everything. Instead, the goal of this work is to analyze and distill important lessons from tech worker organizing and present those findings to the algorithmic fairness research community. As an analogy, consider an iceberg: the tip of it is visible to people, but the underlying structures beneath the surface are not apparent.

We begin by looking at the individual newspaper articles in the archive, which correspond to the tip of iceberg above water. A few of the actions and campaigns were high-profile enough for people to have heard about, but most actions did not receive as much coverage.

Next, we use Sharp's framework to categorize the landscape of actions that are happening. Seeing the scope and nature of the actions being taken has a normalizing effect: when workers see other workers in similar roles organize and win, it shows that such actions are available to them too. Using Sharp's framework, we can identify which tactics are frequently employed (e.g., petitions) and which tactics might still be worth exploring (e.g., majority strikes, noncooperation such as go-slows, etc). By drawing attention to this "negative space" of (so far) less frequently used methods, Sharp's framework provides broader visibility into the landscape beyond what is immediately in front of our eyes.

One challenge in the landscape analysis is that this archive shows tech worker collective *actions*, not campaigns. This means that each entry will, at best, contextualize the action in the previous efforts at the time of writing, but the current structure would not be able to show how a particular action contributed to the success or failure of the overall campaign. Just as campaign *wins* can arrive on a long time horizon, similarly *backsliding* on progress may occur in future years, for instance, if the campaign loses momentum or the company fires the lead organizers. This makes it difficult to readily evaluate the effectiveness of a given action.

Deeper still, we use McAlevey's framework to see how organizers build that power outside of the public eye in general. Whereas mobilizing involves taking visible actions with large groups of people, organizing involves a lot of behind-the-scenes 1-on-1 conversations. Such work is hard to convey in articles, and is often left out of coverage. However, failing to understand that work leaves out a critical component for understanding why some campaigns succeed and others fail.

Finally, not included in this particular project, would be interviewing the campaign organizers themselves to understand their strategy and decision-making. Because this project makes use of public articles and academic frameworks, we "impute" the kind of organizing work that is required for successful campaigns into the "missing data" from the archive. In reality, no campaign plays out exactly the way its leaders expect; the most informative lessons learned usually come from ethnographic and autobiographical works examining the decisions that the campaign leaders made [27, 178]. Further, the reported articles do not have standardized descriptions (e.g., expected goal, number of participants in an action, etc). It will take important followup work in order to get that next level of visibility into the various campaigns.

Chapter 6

Analyzing Differential Privacy

In this chapter, I investigate differential privacy's ability (or lack thereof) to mitigate traditional tradeoffs in data sharing. By generating synthetic data with varying privacy budgets, I perform an empirical investigation into how well differential privacy is able to improve data sharing practices. I then reflect on how this work both reinforces and challenges binary modelling decisions.

6.1 Introduction

The convergence of newly available data, machine learning algorithms, and improved computing resources has greatly increased the efficiency of technical methods to address the large challenges in healthcare. Data scientists have access to computing resources and machine learning algorithms, but the final ingredient for research is data (e.g., from hospitals, insurers, pharmaceutical companies, etc).

In 1996, Congress passed the Health Insurance Portability and Accountability Act (HIPAA) to govern the flow of healthcare information and regulate how individually-identifiable information must be protected by covered entities like hospitals and health insurers. In 2003, the U.S. Department of Health and Human Services (HHS) released the HIPAA Privacy Rule, which prohibited the use and sharing of Protected Health Information (PHI) except for the specific reasons enumerated, such as with a patient's consent or in lieu of that to facilitate treatment, payment, health care operations, or law enforcement requests. Typically, covered entities can only share data with researchers once PHI has been removed, which can be costly.

To understand the potential harms of privacy leaks and data breaches, imagine the worst case scenario. Suppose a hospital's electronic health record (EHR) was posted on the internet for everyone to access without any password requirements. Patients with mental health-related disorders could be discriminated against in ways that are hard to legally contest (e.g., a potential employer who decides to not return an applicant's call). Similar behavior, including interpersonal conflict, could happen if a patient is pregnant or HIV positive. Private notes written by caregivers could reveal information about sexual activity and sexual orientation. And the insurance forms likely contain a patient's phone number, address, and possibly social security number. Access to such information could result in harassment and identity theft. And because this information is *not* available to everyone, there is also the risk that a hacker who obtains unauthorized access to this information could use it as leverage to extort money from the hospital or patients. In other words, the stakes of privacy and security are understandably quite high.

There are great harms that could come to patients if sensitive data is leaked, either directly or through reconstruction [94] and re-identification attacks [76]. Best practices in data sharing (e.g., ethics training, guardrails in the platform interface, secure cloud-based computing, etc.) can mitigate many of these risks, but no approach is ever 100% risk-free. One hope that some have shared is that technical methods might further help sidestep some of these tradeoffs, such as by generating synthetic data which has the same kinds of patterns as in the shared data but without any of the specific individually-identifiable information.

Because of the potential for leakage of training data in generative models, researchers have explored whether Differential Privacy (DP) could ameliorate the issue [271]. Differential Privacy is a technique to add a calibrated amount of noise to the query of a dataset such that it is indistinguishable whether the dataset did or didn't contain any one datapoint [71]. DP provides a "privacy budget" where small budgets give strong privacy guarantees by adding large amounts of noise. In contrast, a large budget would mean a small amount of noise (more accurate) but a necessarily larger risk of leakage. Ideally, scientists would like to be able to create a synthetic dataset which has perfect utility (e.g., a model trained on the synthetic data performs just as well as a model trained on the real PHI-laden data) and negligible privacy risk (e.g., the model is resistant to reconstruction and re-identification attacks).

We build upon prior work to quantify how effectively differential privacy is able to maintain data utility while decreasing privacy risk. Although Stadler, Oprisanu, and Troncoso [271] demonstrate that using differential privacy does not yield models which are both high-utility and low-risk, they do so by measuring risk based on whether they have a large privacy budget or a small privacy budget. Because differential privacy is just one approach for generating hopefully-"safe" synthetic data, such a study does not allow for comparisons against other non-differentially private approaches for generating data to share. In this work, we define measurable notions of privacy, other than the privacy budget, which we then use to measure whether a given dataset is at higher or lower risk for leaking sensitive information.

Our contributions are as follows:

- We review the literature of privacy risks in data sharing.
- We quantify multiple (though by no means all) notions of a dataset's privacy risk in a way independent of how the dataset was generated.
- We generate differentially private datasets (for many values of privacy budget) and characterize the utility-privacy tradeoff curves.
- We conclude that differential privacy continues to not achieve the "sweet spot" of high-utility, low-risk models that some have hoped for.

6.2 Related Work

As interest in data sharing has grown in recent decades, so too have efforts to demonstrate the failures of deidentification techniques.

In 1997, Sweeney was able to identify Massachusetts Governor Weld's medical records in an anonymized dataset by matching his birth date, gender, and zip code [7].

In later work, Sweeney proposed k-anonymity as a method to improve privacy protections [279]. However, k-anonymity is vulnerable to homogeneity attacks (i.e., if the attacker narrows it down to 5 patients, and they all have cancer, then they know the patient has cancer) and background knowledge attacks (i.e., the attacker narrows it down to 5 patients but only one of them is listed to have blue eyes, which the attacker knows) [175].

In their landmark paper, Narayanan et al. 2008 demonstrate that records from the anonymized dataset for the Netflix Prize challenge could be reidentified by linking against public IMDB reviews [193]. Extending this work further, Datta et al. 2012 prove theorems about a variant of the Narayanan-Shmatikov algorithm, and demonstrate suceptibility to isolation attacks and information amplification attacks [66]. Over a decade after their original Netflix Prize attack, Narayanan et al. 2019 reflect that "the core technical insight goes back at least 60 years: a small number of data points about an individual, none of which are uniquely identifying, are collectively equivalent to an identifier" [194].

Generating synthetic data is a generalization of de-identifying data; the goal is to create a dataset without PHI which can be shared. De-identifying the PHI is one of the simplest approaches, though by maintaining the original records, it leaves the dataset much more vulnerable to a linkage attack. The canonical way to create synthetic data is by training a generative model on the sensitive data and then sampling from the generative model to obtain synthetic records. The generative model could be simple (e.g., n-gram counts [146]) or complex (e.g., Generative Adversarial Networks [59]). Complex models involving deep learning or GANs have become powerful enough to generate realistic samples, though sometimes by copying records from the training data [50]. Models might leak information about the training data in a few ways, including:

1. Shokri et al. [261] attack models to find which patients they were trained on. This could reveal sensitive information if the training cohort was constructed using a sensitive attribute (e.g., a model trained on HIV patients).

2. Izzo et al. [128] find that models trained after-the-fact to forget information about sensitive attributes often fail to achieve this as well as intended .

3. Webster et al. [302] finds that some GANs generate images copied from the training set. Generating a real record in its entirety would present obvious privacy harms to that patient.

Even without individual reidentification, there can still be system-level harms, such as when Strava published anonymized walking routes, it showed the location of secret military bases [118].

6.3 Methodology

In this work, we build upon prior work to quantify how effectively differential privacy is able to maintain data utility while decreasing privacy risk. Specifically, we will construct a real dataset, D_R, which will be used to train a model, M_R for a given task. Then, we will train a generative model, $G(D;\epsilon)$ based on summary statistics of a dataset D, with DP privacy budget ϵ. For various values of epsilon, $\epsilon_1, ...\epsilon_n$, we will build synthetic datasets, $D_{S1}, ...D_{Sn}$, by sampling from the generative models $G(D_R, \epsilon_1), ...G(D_R, \epsilon_n)$. We will then use multiple metrics to quantify the privacy risk of the real dataset (D_R) and the synthetic datasets ($D_{S1}, ...D_{Sn}$). Finally, we will then train models $M_{S1}, ...M_{Sn}$ from the synthetic datasets and evaluate the performance of these models.

The result of this procedure will be to create many synthetic datasets for which we can compare their utility and risk for privacy leakage.

6.3.1 Data

To construct the real dataset, D_R, we use the MIMIC-III dataset [132], which contains EHR data of patients who visited the ICU of Beth Israel Deaconess Medical Center from 2001–2012. We use the first hospital admission for 32,660 adult patients. We filter out the 14 patients whose discharge time is reported to have taken place before their admission time.

For each patient, we obtain:

- age (in years),
- length of stay,
- SOFA score [136],
- OASIS score [74],
- race (Asian / Black / Hispanic / Native American / White / unlisted)
- gender (Male / Female),
- admission location (Clinical Referral or Premature Delivery / Emergency Room / Physician Referral or Normal Delivery / Transfer from Hospital / unlisted),
- marital status (Married / Single / unlisted),
- whether patient is a recipient of Medicaid,
- some ICD-9 ontology groupings,[1] and
- some Elixhauser Comorbidity [77] groupings.[2]

Categorical variables were encoded with one-hots. Additionally, in order to quantify the risk to a patient's privacy, we collect a small number of "sensitive attributes," including:

- whether the patient suffers from alcohol abuse;
- whether the patient suffers from drug abuse;
- whether the patient is HIV positive; and
- whether the patient has cancer;

In this work, we treat nonsensitive attributes as vectors of attack from adversaries with linkable datasets. In other words, we consider attacks along the lines of "suppose I could re-identify someone based on their age, marital status, etc. Would pairing that ability with this dataset allow me to conclude whether they are HIV positive?"

This real dataset is split 70/30 into train/test sets, $D_{R_{train}}$ and $D_{R_{test}}$, respectively.

[1] ICD groups for infectious (1-139), neoplasms (140-239), respiratory diseases (460-519), skin diseases (680-709), and injury (800-999).

[2] Elixhauser groups for: weight loss, pulmonary circulation disorder, paralysis, blood loss anemia, chronic pulmonary disease, drug abuse, and alcohol abuse.

The real training set will be used to generate differentially private synthetic datasets, D_{Si}, from. Both $D_{R_{train}}$ and the many D_{Si} datasets will be evaluated for privacy leakage and how much clinical utility they retain. The real test set, $D_{R_{test}}$, will be used to measure the clinical utility of the above datasets, as described in Section 6.4.1.

6.3.2 Generative Model

In this section, we describe how we build a differentially private generative model $G(D; \epsilon)$ by learning summary statistics of a given dataset D and adding noise according to the "privacy budget" ϵ.

Differential privacy is a system for sharing data publicly while maintaining formal guarantees about resistance to membership inference attacks. Differential privacy is not a particular method but is instead a definition: a query Q (e.g., count, sum, etc.) that operates on a dataset is said to be ϵ-differentially private if for all pairs of datasets D_1 and D_2 which differ by only a single element and all events S (e.g., query returns a value of 15):

$$Pr[Q(D_1) \in S] \leq e^{\epsilon} \cdot Pr[Q(D_2) \in S] \tag{6.1}$$

This is best understood with a motivating example: suppose I have a running tally of how old everyone in the classroom is, and then someone 105 years old enters the room. My exact count would increase by 105, and I might be able to reverse engineer who entered the room because of how few people are that exact age. On the other hand, if my running tally were differentially private, then the reported results of the query might not say someone 105 entered the room; noise is added to the aggregate count which obscures the exact contribution of any one participating individual. We could add a large amount of noise (at the cost of accuracy) or a small amount of noise (at the cost of privacy), and this knob is determined by the privacy budget ϵ.

Differential privacy is closed under compositionality. Specifically, there are rules for sequential compositionality (e.g., querying average age of everyone, and then querying average age of everyone not named William) and parallel compositionality (i.e., performing a query over disjoint groups, such as querying the number of people below 30 years old and querying the number of people above 30 years old). For sequential compositionality, the output of performing an ϵ_1-differentially private query and an ϵ_2-differentially private query is also differentially private, with a budget of $\epsilon_1 + \epsilon_2$ [71]. On the other hand, parallel compositionality allows for combining disjoint queries "costlessly," where the output of performing an ϵ_1-differentially private query and a disjoint ϵ_2-differentially private query is also differentially private, with a budget of $\max(\epsilon_1, \epsilon_2)$. This is often used to publish differentially private histograms with low cost to the privacy budget [310].

Researchers have also discovered methods used for adding differential privacy into the machine learning training process, using techniques such as Differentially Private Stoachstic Gradient Descent (DP-SGD) [236] and Private Aggregation of Teacher Ensembles (PATE) [219]. IBM developed a public differential privacy library, fittingly named "DiffPrivLib" [120]. It uses the "Smooth Sensitivity" method to train the random forest classifier [86].

Building the Generative Model

To build the generative model, $G(D; \epsilon)$, we partition the feature space into three sections:

1. demographic binary features (13 dim);
2. ICD-based binary features (13 dim); and
3. continuous features (age, OASIS, SOFA, length of stay).

We build six separate generative models, two count-based models (one for each of sections 1 and 2) and four diffprivlab-based models (one per continuous feature).

Following the structure of the original data, we enforce constraints that a given patient can have at most one race, marital status, admission location, and gender. This is done by collapsing these variables into non-binary categorical variables before counting co-occurrences. To avoid the exponential "blow up" of 2^{26} combinations (or even the $\approx 2^{20}$ combinations after collapsing demographic variables to be non-binary), we make a simplifying assumption that demographic variables are independent from ICD-based variables. Although such an assumption is not strictly true, it allows us to decompose the task into learning significantly smaller generative models (360 combinations and 8192 combinations) and concatenating the results.

We construct two histograms, one per section, S, to determine exact counts for how many patients contain each combination of features. Because of parallel compositionality, partitioning a dataset into k mutually exclusive sets before performing the query only costs one charge to the privacy budget instead of k charges [71]. We add noise using the Laplace mechanism to each section with privacy budget ϵ_S. For an overall privacy budget of ϵ for $G(D; \epsilon)$, both histograms are allocated a budget of $\epsilon_S = \frac{\epsilon}{6}$. This amounts to adding Laplace(0, $\frac{\epsilon}{6}$), independently sampled, to each cell of each histogram.

For the age, SOFA, and OASIS features — which are non-binary — we use IBM's diffprivlib library to learn a differentially private random forest to predict those three attributes from the 26 earlier binary features. Each random forest has 100 estimators, a max depth of 3, and a privacy budget of $\frac{\epsilon}{6}$. In order to ensure the generative model is stochastic, we perform classification instead of regression; we discretize each target into ten equally-sized bins and then learn to predict which bin is associated with the features. This allows us to select the appropriate bin and then stochastically select that feature's value uniformly from the bin. Differential privacy requires specifying the maximum value (rather than inferring it from the data, which could leak information), so we clip all age values to 90 years old, SOFA scores to 22, and OASIS scores to 70.

The final attribute, length of stay, is also a non-binary variable whose value we generate with a diffprivlib random forest. One difference from above is that this model uses a 29-dimensional feature vector, because age, SOFA, and OASIS values are concatenated to the earlier 26-dimensional vector.

We can sample from $G(D; \epsilon)$ by:

1. sampling 13 dimensions from the demographic DP-histogram;
2. sampling 13 dimensions from the ICD-based DP-histogram;
3. concatenating the demographic and ICD-based features into a 26-dimensional

binary vector;

4. predicting which age, SOFA, and OASIS bins are most associated with that feature vector using their respective differentially private random forests;
5. randomly sampling values from each selected bin;
6. concatenating the age, SOFA, and OASIS scores onto the demographic and ICD-based features into a 29-dimensional feature vector;
7. predicting which length of stay bin is most associated with that 29-dimensional feature vector using the differentially private length-of-stay random forest; and
8. randomly sampling a length of stay from the selected bin.

This procedure is run 200 times to generate 200 synthetic datasets. Using the definitions of utility and privacy risk defined in Section 6.4, each dataset is evaluated for its utility (i.e., how well a model trained on it performs on the heldout real test data) and its privacy (i.e., four Membership Distance metrics and four Attribute Inference metrics). We construct eight utility-privacy curves, each one showing the utility and privacy of each of the 200 synthetic datasets.

By the sequential compositional property of differential privacy, the individual privacy budgets for each procedure sum to an overall privacy budget of:

$$\frac{\epsilon}{6} + \frac{\epsilon}{6} + \frac{\epsilon}{6} + \frac{\epsilon}{6} + \frac{\epsilon}{6} + \frac{\epsilon}{6} = \epsilon \tag{6.2}$$

6.4 Measuring Utility and Risk

In this section, we describe how to evaluate the clinical utility and privacy risk of a datset, whether it is the real data $D_{R_{train}}$ or synthetic data, D_{Si}.

6.4.1 Utility: Model Performance

For a given dataset D, we build a model, M_D, to predict the patient's length of stay. We use an XGBoost regressor [56] model with 100 estimators and a max depth of 6. We then evaluate this model by measuring the absolute error when trying to predict the lengths of stay from patients in the real test set, $D_{R_{test}}$.

6.4.2 Privacy: Membership Distance

As prior work has shown, sparse, high-dimensional data is often at high risk for privacy attacks [193]. In high dimensions, data subjects are vulnerable to being singled out [63]. Even in approximated scenarios [279], adversaries can learn information if they can narrow their search down to a smaller number of candidates.

For intuition, suppose the real dataset had an 83-year-old Native American married woman who is HIV positive. Further, suppose a dataset is shared which contains an 82-year-old Native American married woman. Even though the age is not an exact match, an adversary might conclude (if the syntethic data was derived from real data) that the public "synthetic" row accidentally memorized a real row. This membership inference attack would allow the advserary to re-identify the woman and could then perform additional attacks based on linked datasets.

For this privacy metric, we formalize that intuition by trying to identify each real patient's synthetic "doppelganger." To evaluate a dataset, D, we have one metric per sensitive attribute. For each sensitive attribute, A, we first obtain the group of patients who have that sensitive attribute (i.e., the group of patients who have cancer); this set of patients is denoted P_A for the real dataset and F_A for the synthetic dataset. For each patient $p \in P_A$, we find the synthetic patient from F_A that has the highest similarity with p when comparing only the nonsensitive attributes (i.e., its "doppelganger"). Finally, we aggregate across all patients in P_A by taking the median Euclidean distance between each real patient and their doppelganger:

$$PrivDist_A = median_{\mathbf{p} \in P_A}(min_{\mathbf{f} \in F_A} || \mathbf{p}_{nonsensitive} - \mathbf{f}_{nonsensitive} ||) \tag{6.3}$$

An exact match between a real patient and its doppelgangers would have a distance of 0, which would indicate both datasets list someone with attribute A and identical nonsensitive attributes. On the other hand, a high value for this metric would indicate that real patients do not have obvious "doppelgangers" in the public data.

6.4.3 Privacy: Attribute Inference

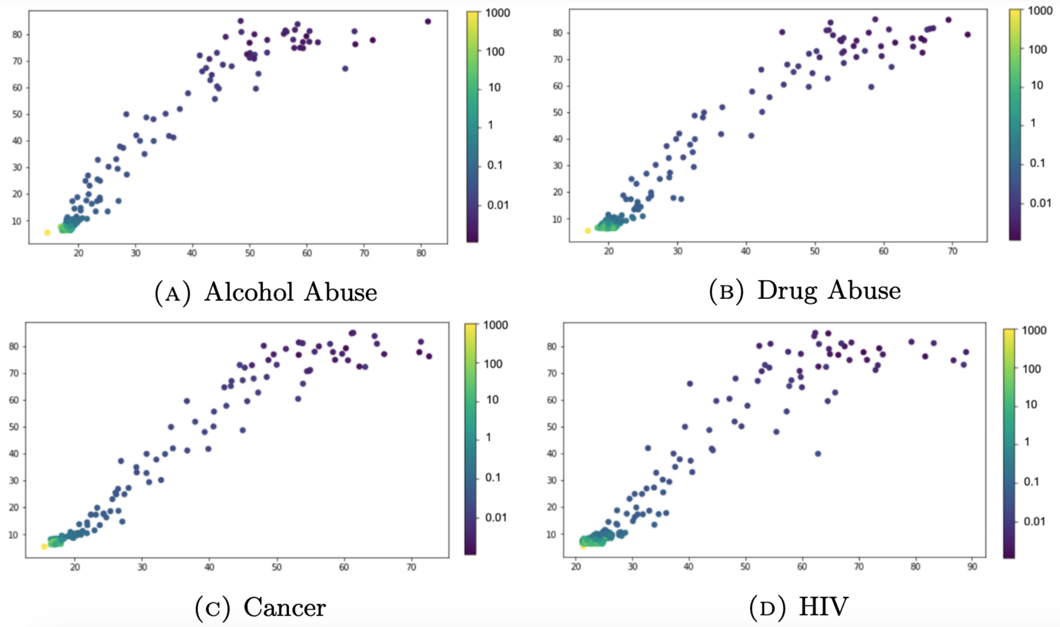
If there are correlations between the sensitive attributes and some of the nonsensitive attributes, an attacker might be able to probabilistically infer that information about patients. Certainly this is part of what makes privacy so difficult to preserve without any tradeoffs; associations between the features and target (length of stay) are considered useful whereas associations between the features and sensitive attributes could be potential privacy leakage.

As above, to evaluate a dataset, D, we have one metric per sensitive attribute. For each sensitive attribute, A, we train a classifier, $C_{A,D}$ to impute whether the patient has the sensitive attribute based upon the sparse nonsensitive features. This classifier is an XGBoost classifier trained with 50 estimators and a max depth of 2. This learns what relationships the shared dataset makes between sensitive attributes and sparse, high-dimensional features that adversaries may have access to. We then compute the recall of identifying patients with that sensitive attribute in the test data $D_{R_{test}}$. To avoid the chance that this is “gamed” by a classifier which always predicts positive, our classifier only predicts n_D patients to be positive, where n_D is the actual number of positive patients in D. Using recall allows for an intuitive understanding of this metric: a score of 0.3 indicates that an adversary which knows the patient’s age, gender, nonsensitive ICD groups, etc. would be able to correctly identify 30% of patients which have sensitive attribute A (e.g., drug abuse).

6.5 Results

In this section, we report the 8 utility-privacy curves for the Membership Distance (4 sensitive attributes) and Attribute Inference (4 sensitive attributes). Additionally, we explore these curves more closely for the *ICD=Cancer* sensitive attribute metrics.

Figure 6-1: **Membership Distance** curves of each sensitive attribute. This notion of utility is defined in Section 6.4.1. This notion of privacy is the Membership Distance definition found in Section 6.4.2. The x-axis is Euclidian distance units and the y-axis is Mean Absolute Error. Each point's color corresponds to that dataset's privacy budget, ϵ.



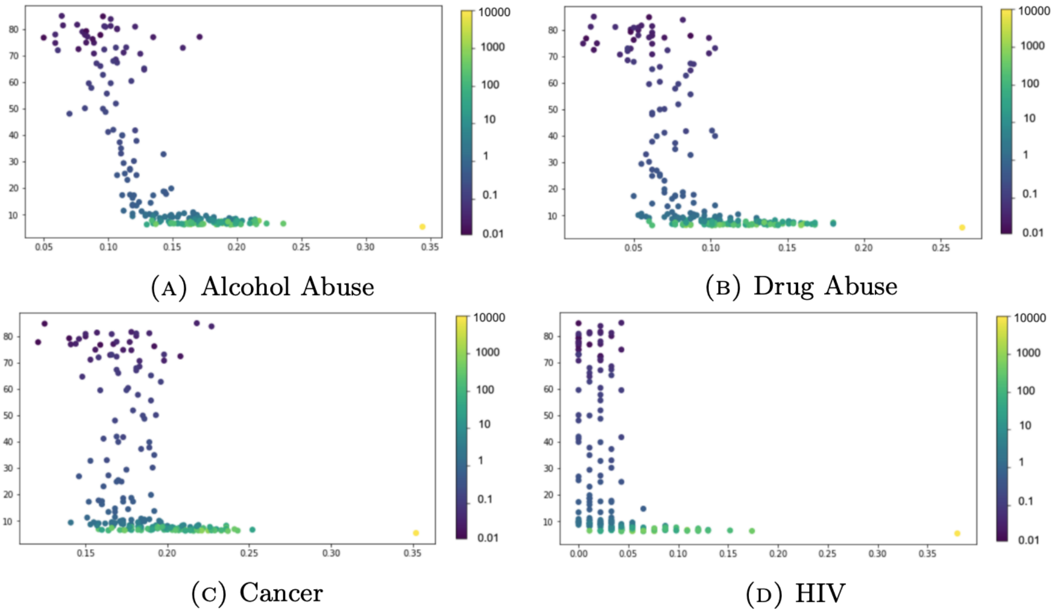
(A) Alcohol Abuse

(B) Drug Abuse

(C) Cancer

(D) HIV

Figure 6-2: **Attribute Inference** curves of for each sensitive attribute. This notion of utility is defined in Section 6.4.1. This notion of privacy is the Attribute Inference definition found in Section 6.4.3. The x-axis is recall from an attribute prediction classifier and the y-axis is Mean Absolute Error. Each point's color corresponds from the privacy budget, ϵ, used to generate that dataset.



(A) Alcohol Abuse

(B) Drug Abuse

(C) Cancer

(D) HIV

6.5.1 Utility-Privacy Curves

Membership Distance

Figure 6-1 shows the utility-distance curves for the Membership Distance definition of privacy from Section 6.4.2.

The yellow point in each figure corresponds to the real dataset. The real data was not generated via the generative model described in Section 6.3.2, which causes it to have a discontinuity from the rest of the curve for most sensitive attributes. This is because the generative model has non-reversible encodings, such as the the continuous variables (age, SOFA, OASIS, length of stay) being encoded and decoded at the granularity of "which of the ten bins does this fall into?" Even without any differential privacy noise, such a process does not allow perfect reconstruction.

All four curves have similar trends, which we will explore in more depth in Section 6.5.2. There is a near-proportional relationship between the privacy metric and the model performance. The metric measures the median distance between patients with that attribute and their closest "doppelganger" with that attribute. The larger the median distance is, the more difficult it is to "single out" patients who have the sensitive attribute.

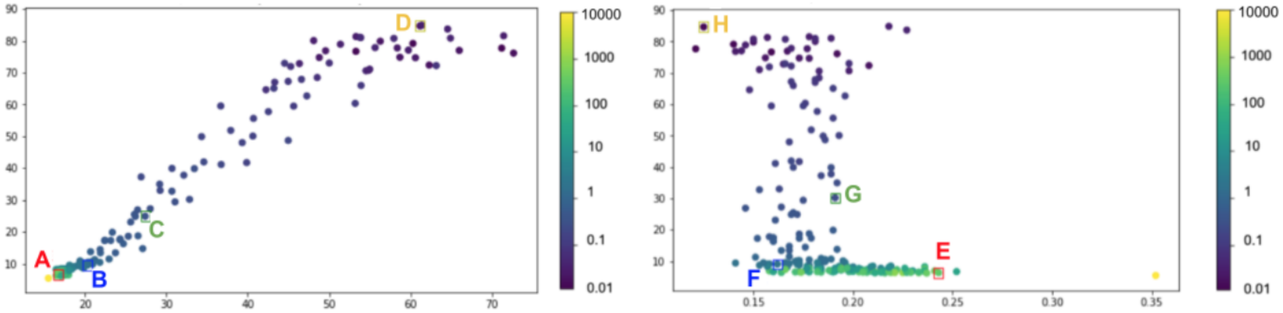
Because of the strong correlation between error rate and difficulty-from-singling-out, this generative model demonstrates the same utility vs. privacy tradeoff that plagues traditional synthetic data. Even though differential privacy has mathematical guarantees about the kind of noise it is adds, those properties do not allow the generative model to produce low-risk models that are able to retain clinical utility.

Attribute Inference

Figure 6-2 shows the utility curves for the Attribute Inference definition of privacy from Section 6.4.2.

The discontinuity between the real (yellow) point and all synthetic points is even more stark than in Figure 6-1. This likely is due to the independence assumptions in the generative model, where some attributes were assumed to be independent from

Figure 6-3: Restating the Membership Distance (A) and Attribute Inference (B) curves for the cancer metric. Each curve is labeled with four letters (A-D, E-H) where each letter corresponds to a dataset visualized in Figure 6-5.



(A) Membership Difference

(B) Attribute Inference

each other (and then were combined to generate the target variable in a differentially private way). Modelling assumptions — most notably that ICD-based features and demographic features were assumed to be independent — severed relationships between attributes which likely were useful for attribute inference. As a result, every single synthetic dataset makes it much harder than the real data to infer sensitive attributes based on nonsensitive attributes.

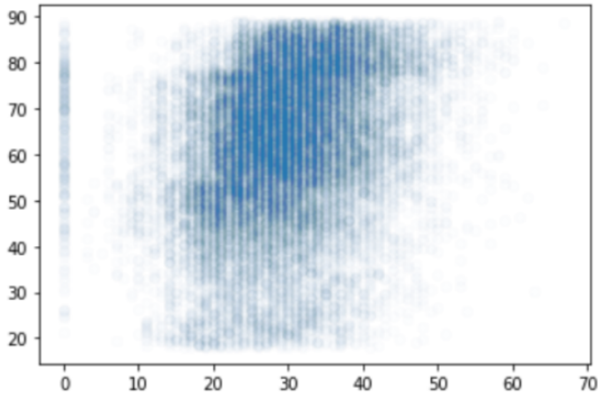
6.5.2 Closer Look: Privacy for Cancer Patients

In this section, we analyze the utility-risk curves a little more closely to better understand what kinds of datasets are being generated. Figure 6-3 shows the Membership Distance and Attribute Inference curves for the cancer-based privacy metrics. These figures identify representative points along their curves, labeled A-H, which we will examine further.

In order to visualize these datasets, they must be projected from 30 dimensions down to 2 dimensions. Because the data has a mix of sparse, high-dimensional data (26 dimensions) and dense, continuous variables (4 dimensions), methods like t-SNE and PCA are liable to distort the direct values. Instead, we visualize the data by projecting onto 2 of the dense dimensions: age and OASIS. Figure 6-4 shows the relationship between these two attributes in the real dataset.

When comparing the real data against the projections in Figure 6-5, we see that all

Figure 6-4: 2-dimensional visualizations of the real dataset. This view is a projection of the OASIS (x-axis) and age (y-axis) dimensions.

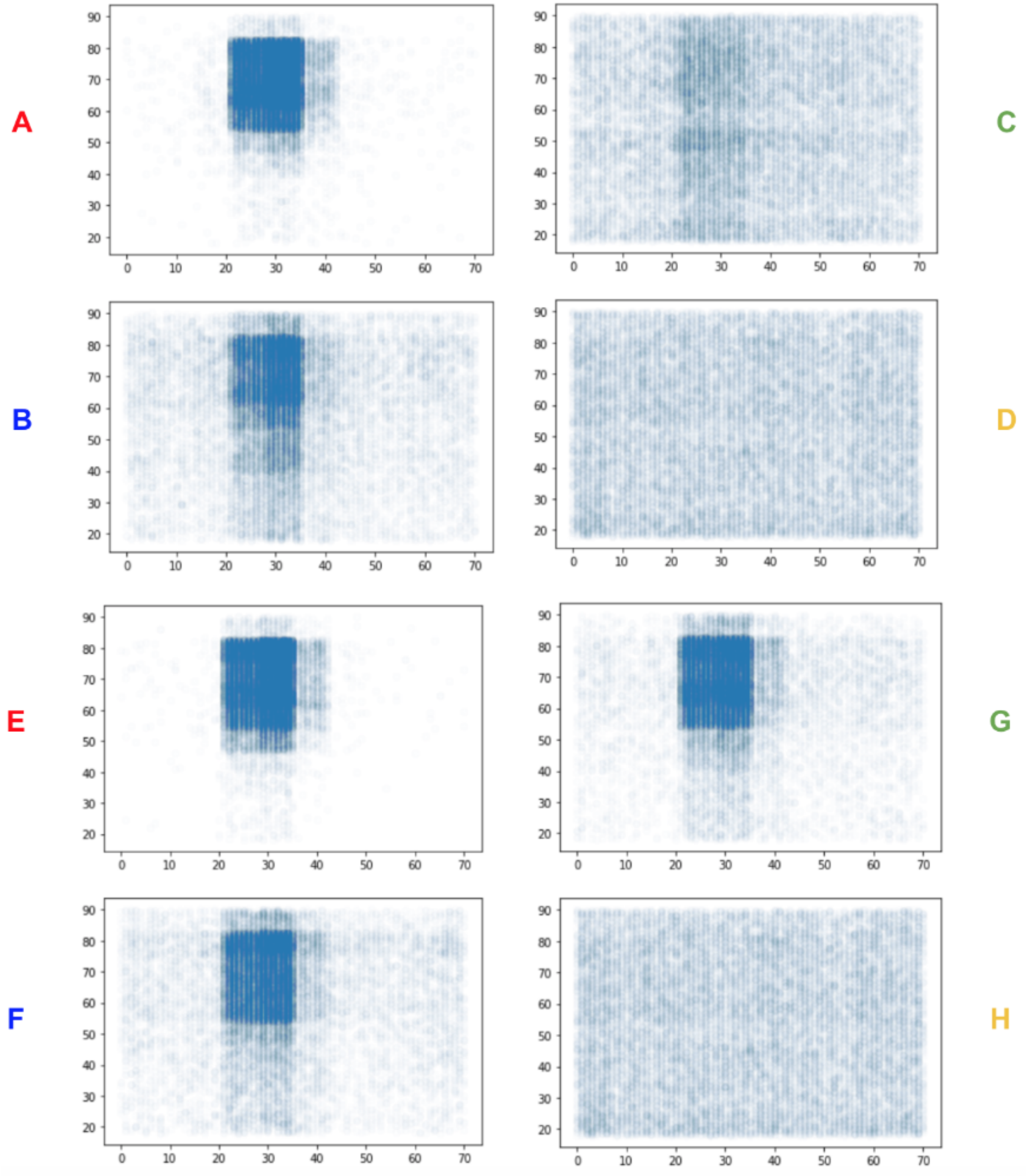


of these generative models are limited by the bin-based generative model for producing age and OASIS. Both dimensions are generated via differentially private random forests which predict which of ten bins to sample from. But once a bin is selected, the sampling is done over a uniform distribution in that range. As a result, the generative model "throws away" the fine-grained relationship between variables. Although we can see in the real data a modest positive correlation between age and OASIS, the synthetic datasets largely indicate no correlation between the two dimensions.

Points A, B, E, and F achieve relatively good clinical utility in Figure 6-3. Their 2d projections have much less noise than points C, D, and H. Although these higher utility points are constrained by the generative model bin-based structure, they nonetheless do concentrate probability mass in the appropriate region of the space. Points C, D, and H add too much noise to the point where all signal is lost.

On the whole, this demonstrates the noise mechanism at work: adding a lot of noise (i.e., a small privacy budget) will distort the relationships between variables that are able to be captured. Although, yes, it does lead to higher levels of privacy, it achieves this by degrading the meaningful signals as well.

Figure 6-5: 2-dimensional visualizations of the 8 datasets identified in Figure 6-3. For comparison with Figure 6-4, the OASIS score is the x-axis and the age is the y-axis.



6.6 Discussion

A benefit of differential privacy is that it allows the user to control how much "privacy leakage" they can allow from their data. By adding more noise or less noise, the user is able to guarantee bounds on whether an adversary could determine whether an individual contributed their data. However, that alone is not enough to sidestep the difficult issues in data sharing.

Even aside from differential privacy techniques, the so-called "Holy Grail" that researchers are hoping to develop is an anonymization mechanism which removes all privacy risk without impacting the data's utility. However, as we saw, the line between utility and privacy is not always as clearly defined. Certainly medical conditions such as HIV status and pregnancy are considered sensitive by most people, however, other attributes such as marital status or even gender may be very sensitive to others. Although there is a desire for technical experts to design a "scrubber" which acts as a "one stop shop," the task of ensuring privacy is quite challenging. Arvind Narayanan has been critical of efforts which hope to deidentify data sufficiently to be able to release for public use with sufficient privacy protections [195].

Differential privacy is a useful tool in the right situations, but policymakers should be aware of what it can do as well as what it cannot do. One open challenge that policymakers have struggled with is what is the appropriate way to set their privacy budget; the unit-less ϵ variable is useful in the definition of differential privacy, but the current state of research offers little practical advice of how to find the right tradeoff.

For the most part, selecting the tradeoff point is not a technical issue. Instead, researchers must work with knowledgeable domain experts to map the privacy interests and the threat models to defend against. But to the extent that technical methods can facilitate the process, researchers should further develop approaches for measuring privacy risk. Much like the past 6 years have seen a plethora of mathematical definitions of fairness [186], so too could experts model different quantifiable notions. Then the right tool could be selected for a given context (e.g., by presenting a few options to domain experts and discussing the important areas of focus).

This work presents contributions in that direction by bypassing the privacy budget when evaluating differentially private methods. Because stakeholders and policymakers will have no intuition for how to answer a question like "Is a budget of ϵ=0.8 sufficient to keep these records safe enough?" it is important that we move to interpretable notions.

6.7 Project "Evidence"

In this work, I moved beyond a privacy budget-driven notion of privacy loss to quantify the utility-privacy tradeoff curve offered by a case study in differential privacy. Just as we use principles of Data Feminism to reflect on the Tech Worker Organizing project in Chapter 5, so to can we use another principle to interrogate the assumptions of this work. The fourth principle of Data Feminism is to "rethink binaries and hierarchies." This work both made its own reductive, binary modelling assumptions (e.g., sensitive variables) and also directly challenged the notion that differential privacy, itself, is a binary design choice (as opposed to a spectrum, whose guarantees of privacy are highly dependent on the amount of noise).

6.7.1 Spectrum of Differential Privacy

Since differential privacy was created in 2006, it has garnered a lot of attention as a tractable way to quantify the degree to which sensitive information could be shared. Multiple efforts to employ differential privacy attracted initial praise from privacy advocates, but as their implementation rolled out, numerous challenges emerged. One such challenge was the concern that using extraordinarily large privacy budgets effectively nullified any benefits from theoretical guarantees and likely led to data whose quality has been lowered without a provable benefit in practical privacy protections. I begin with an overview of two such examples and then discuss this work's contribution to whether a project claiming it "uses differential privacy" is meaningful in its own right.

2020 US Census. Every ten years, the US Census enumerates all people living in the United States and releases aggregate statistics, which are used for many purposes: determining proportions for Congressional representation, determining whether Congressional districts comply with the Voting Rights Act, allocating funding resources, and more. With the sharp increase in computational resources, officers at the Census Bureau worried that despite deidentification methods, a large number of respondents could be reconstructed from aggregated summary statistics [94]. In response to these concerns, the Bureau released the 2020 Census using differential privacy to add noise to the counts. As documented by danah boyd [94], this process led to numerous challenges and growing pains. In addition to critics who argue that deliberately adding noise to census data is bad policy with little-to-no measurable upside [318], others worry that the technology of differential privacy is too immature to use effectively. For instance, the Bureau chose to add less noise (i.e., larger privacy budget cost) on the "spine" of the counting process [39], to ensure the most important information for Congressional apportionment was not too far from reality, but its not clear how to "strike the right balance." The overall Census reporting totaled a privacy budget of $\epsilon = 19.61$. Is that too much noise? Too little noise? Is the noise correctly distributed to prioritize the right accounting? Mathematically speaking, a privacy budget that large suggests a nearly meaningless privacy guarantee: the ratio of probabilities for how much one element's exclusion could change the reported data is upper bounded by a factor of $e^{19.61} = 328,484,431$.

Apple. Whereas the Census Bureau obtained an exact count and then used differential privacy to add noise to the publicly released aggregated statistics, Apple also used differential privacy but did so in a decentralized manner. In 2016, Apple announced they would be using local differential privacy to collect data about their users and sending noisy values to their centralized servers for analysis [281]. Although the differential privacy research community applauded the effort, many offered critiques on the execution, including Frank McSherry — one of the inventors of differential privacy — who said that "Apple has put some kind of handcuffs on in how they interact with your data. It just turns out those handcuffs are made out of tissue paper" [104]. The researchers who reverse engineered Apple's code found a privacy budget as high as $epsilon = 43$ under some settings, but McSherry commented that "any [value of epsilon] much bigger than one is not a very reassuring guarantee. Using an epsilon value of 14 per day strikes me as relatively pointless." Much like with the census, experts are skeptical that merely using "differential privacy" added meaningful protections to the data.

In this work, I create a differentially private model and explore 200 sampling runs with different values of privacy budget. Doing this allowed me to trace the "utility-privacy" curve(s). Notably, one major observation that we saw was that differential privacy allowed for a relatively graceful transition between good utility and good privacy. On the one hand, the use of differential privacy was disappointingly unable to sidestep the traditional tradeoffs between two values in tension with each other. On the other hand, having this mechanism does allow for more fine-grained control by the data producers to determine the quality of published data depending on their risk tolerance. However, more work is required for publishers to determine the right privacy metrics for their particular task and then they must effectively engage with stakeholders to understand where on the tradeoff curve could be sufficient. Historically, such fine-grained control has not been available in this way, which means there will likely be mistakes and growing pains if differential privacy continues to be used for high-profile data publishing.

Of course, it may also prove to be the case that negative receptions of differential privacy case studies may deter future data publishers from adopting the technology; to date, it has been seen as a binary design improvement that can earn a cycle of good publicity, though perhaps that may change.

6.7.2 (Multi-)Spectrum of Data Sensitivity

In Section 6.3.1, I introduced a distinction between "sensitive" attributes and non-sensitive attributes. This binary sensitive/nonsensitive paradigm is a simplifying assumption, both because there is a spectrum of how sensitive information is (e.g., right handedness is less sensitive than weight, which is likely less sensitive than a family history of substance abuse) and also because different people have different notions of which attributes are most sensitive for themselves. Such value determinations must be taken into account when modeling the risks and threats in order to understand what is trying to be protected by a notion of privacy.

Because of the diversity in values, we would require the addition of other principles of Data Feminism — including "Examine Power" and "Embrace Pluralism" — to fully characterize how different forms of oppression act on the data subject based on their intersecting identities. For instance, after Narayanan et al. 2008 demonstrated they could reidentify users in the anonymized Netflix prize challenge dataset [193], a closeted lesbian woman sued Netflix because she could be outed if people saw her history of watching LGBT films [264]. Similarly, although many data subjects would consider "gender" to be not particularly sensitive, that is far from true for everyone; transgender users of an online platform might not be out to all of their online friends (e.g., to family back home).

In this project, I make modeling assumptions about which features are sensitive, but more work must be done to explore how to account for the distribution of concerns. For example, one notion may involve focusing on consensus areas of sensitivity whereas a different notion may measure progress by how well the modeling performs for the worst-case. The complex nature of diverging values is in tension with the fine-grained level of control that differential privacy affords to data publishers. It will be essential to analyze case studies with specific threat models and examine how effectively each one was able to account for protecting the attributes that its data subjects are concerned about.

Chapter 7

Conclusion

Although hundreds of organizations are creating top-down statements of principles and values, such efforts are hampered by vagueness. Of course everyone agrees that a process should be “fair,” but we do not have the policies or norms in place to adjudicate whether a given thing is or isn’t fair as issues arise. Instead, I propose we employ Evidence-based AI Ethics: we should learn bottom-up from case studies. In Chapters 2–6, I demonstrate how to use critical lenses (e.g., Effective Altruism and Data Feminism) to extract “evidence” from projects. In this final chapter, I sketch the crucial next steps: how institutions can learn from this evidence and put relevant policies and norms into place.

The rest of this chapter is as follows: in Section 7.1 I describe many of the relevant actors in the AI Ethics ecosystem and their incentives, and in Section 7.2 I demonstrate two examples of efforts to build norms and advocate for policies informed by case studies.

7.1 Who is Doing AI Ethics?

In Chapter 1, I discuss the role and incentives that academics face. In this section, I overview the roles and incentives of other actors that contribute to AI Ethics.

7.1.1 Corporations

Corporations — particularly "Big Tech" companies like Facebook, Google, Microsoft, etc. — play a large role in the norms that are created for developers and engineers. Whereas academic institutions canonically do basic research, corporations are intended to translate that research into products and services that are valuable enough to people that customers will buy them. In simple models, long-term incentives of rational actors would compel corporations to always act ethically, but other factors (e.g., information asymmetries, cognitive biases, calculated risk of getting caught, cost-benefit tradeoffs) can cause them to deviate from those simple models. As such, most of the other actors in the ecosystem sometimes monitor corporations to hold them accountable.

As mentioned in Chapter 1, there are dozens of corporate principles of AI Ethics. However, it's not clear how meaningful or effective these guidelines are in general, and there is likely large variation from firm to firm. Using Google as an example, Google created its *AI Principles*[1] in response to employee activism to protest the company's contract with the Department of Defense for drone technology. The firm also created a review process[2] to translate these principles into an assessment of each product, and (according to Google) the assessment convinced them "to hold off on offering functionality before working through important technology and policy questions" for general-purpose facial recognition API. However, others contend that the current assessment process is insufficient, citing examples such as how its dermatology app was not tested sufficiently on darker skin [83]. Additionally, after the contentious departures of both leads of Google's Ethical AI team, the rest of that team is raising concerns about the insufficiency of internal processes [100].

7.1.2 AI Workers

Chapter 5 discussed the role of tech workers in AI Ethics and oversight in great detail. In particular, workers can play a very important role in holding corporations

[1] https://blog.google/technology/ai/ai-principles

[2] https://ai.google/responsibilities/review-process

accountable through a large spectrum of tactics.

7.1.3 Policymakers and Regulators

Companies often respond to public perception and market forces, but they might not always do that. Sometimes — such as in the case of market failures or when a negative externality would cause harm without sufficient disincentive to the actor — legal power is the best way to address a problem. For instance, the Federal Trade Commission (FTC) has a broad mandate to protect consumers from fraud and deception in the marketplace. If the FTC finds that a company is engaging in unfair or deceptive acts or practices, then they can hold the firm accountable with lawsuits and civil penalties. Likewise, Congress has oversight abilities (backed by subpoena power) to call companies to testify before them as they consider writing new laws.

Ideally, the goal of policymakers is to represent the will of the people and to translate popular opinions into law. In practice, there is a wide distribution of competing incentives among politicians running for re-election. But because this is not a political science thesis, we can use the simpler model that policymakers are trying to pass laws which codify popular will. Other actors, including academics, corporations, think tanks, the public, and more often meet with them to influence the decisions that they make.

Historically, US policymakers have a bad track record of ensuring the law keeps up to date with technological progress. For instance, in 1986, Congress passed the Electronic Communications Privacy Act (ECPA) to extend restrictions on government wire taps of telephone calls to include transmissions of electronic data by computer. However, this law was based on a 1980s understanding of data, where emails were only stored on the owner's computer. To balance interests at the time, Congress concluded that if an email were on someone else's server for 180 days or more, then the data would be considered "abandoned" and the police would no longer need a warrant to access it. This model is no longer striking the intended balancing of privacy interests now that cloud-based web services (e.g. gmail) store *all* emails on a third-party server like Google's. Despite the law being over 30 years old, the "180 day rule" is still technically the law (although in practice, courts have decided to ignore that rule).

7.1.4 Think Tanks and Advocacy Groups

Congress does not have enough in-house technical expertise to be able to keep up with advancements in technology, so other organizations — such as think tanks — have emerged to develop a body of knowledge relevant to specialized interests. These organizations are typically funded by donors or grants, and focus on a specific interest, such as the Electronic Frontier Foundation (EFF), an international non-profit digital rights group which often weighs in to Congress and the Courts about questions involving digital privacy.

Other nonprofits, such as the Center for Humane Technology, focus on raising public awareness about the harms or benefits of new technology. The EFF and American Civil Liberties Union (ACLU)[3] file lawsuits to defend civil liberties like privacy and free speech.

Their incentive is to convince their donors they are sufficiently making a positive impact.

7.1.5 Media and Journalists

The purpose of the media and journalists is to keep the public informed. Much like with policymakers, there are interesting aspects of media organizations' incentives which will be left for others to study [149], but the main takeaway is that media organizations compete for attention of their viewers as the way to financially support themselves.

Investigative journalists and watchdog organizations have played an oversight role in AI Ethics behavior, especially since the US government has mostly taken a "light touch" regulatory approach for the tech industry. One early high-profile discussion of the current wave of AI Ethics discourse came from ProPublica's "Machine Bias"

[3] Disclosure: I interned for the ACLU of Massachusetts in 2021

investigation of a risk assessment tool, showing that Black suspects were more likely than white suspects to be incorrectly given a "high risk" assessment [11]. Although fairness is arguably the most-discussed topic in AI Ethics in 2021, other challenges have also been covered by media, including self-driving cars [110, 38], algorithmically-amplified misinformation [144], and biometric surveillance [265]. Ouchchy, Coin, and Dubljevi [215] read a comprehensive set of 254 media articles written 2013–2018 about AI Ethics, and found that the tone of the articles has become increasingly critical as additional high-profile instances become public, with particular emphases on unintended consequences, accountability, and a lack of ethics.

The media has a large influence on public discourse. Many books aimed at a general audience have spoken about the potential harms of tech, including Algorithms of Oppression [206], Weapons of Math Destruction [208], Automating Inequality [78] Data Feminism [64], Artificial Unintelligence [42], Technology After Race [24], The Ugly Truth [87], and more. Relatedly, numerous documentaries have sought to critique massive data collection [6], targeted and behavioral advertising [213], and facial surveillance [143]. These artifacts allow for broadcasting of concepts and critiques to the public.

7.1.6 The Public

In some sense, the public is the most powerful domain because all of the other domains derive their power and legitimacy through public will. A lot of what the public understands is filtered through the media, and to a certain extent academics, politicians, think tanks, and corporations directly.

Because of the public's influence on all of the other domains, changes in public opinion can have large impacts on how the other agents behave. For instance, since the 2018 reveal of Facebook's scandal involving Cambridge Analytics (and numerous additional scandals since), politicians have been much more openly critical of Facebook, and the company has changed its data sharing practices to try to avoid further scrutiny.

7.2 Building Norms and Policies

There are many different kinds of actors in the space: academia, regulators, media and journalism, think tanks, corporations, and the public. This “alignment” of shared understanding which actions are/aren’t ethical does not just happen on its own. As we have seen from Evidence-based Medicine, there is a lot of meta-organizing work which goes into building the consensus within a field that can promote desirable behaviors (e.g., medical practices likely to help patients, data collection practices which respect subject preferences) and discourage undesirable behaviors (e.g., promoting harmful content online to increase user engagement).

To build this movement of ethical uses of AI, institutions will need do the work of generating consensus:

- e.g. Academics will need to organize more conferences to discuss AI Ethics, particularly around the analysis of and lessons learned from case studies.
- e.g. Journalists will need to continue to serve as watchdogs and do investigative journalism to uncover how existing practices are operating.
- e.g. Academics will need to develop curricula which teach the values.
- e.g. Companies will hopefully compete on ethics in their brand.
- e.g. Regulators and courts will create and iterate on the legal frameworks for governing data practices.

I demonstrate two efforts I’ve contributed to which have worked to build better norms and policies.

7.2.1 Policy Synthesis: AI Policy Forum

The MIT AI Policy Forum (AIPF) is a global effort convened by MIT to formulate concrete guidance for how governments and companies can handle emerging issues that arise in AI. In Spring 2022, AIPF hosted three workshops where diverse groups of stakeholders discussed policy implications of AI in healthcare, finance, and mobility/transportation.

I worked with the healthcare workshop organizers. The event was across two days in January 2022. The AIPF convened a group of AI health policy experts from academia, the non-profit sector, government and commercial services, spanning disciplines including computer science, medicine, law, and anthropology. Twenty-four speakers presented their work: deploying AI in hospitals, creating data sharing policies for nonprofits and governments, using data to inform Greece's COVID testing allocation, technical assessments of data sharing tools, ethical frameworks for machine learning, and more.

The result of the discussion was a report [31] which made a series of policy recommendations based on specific bright spots and challenges identified by the case studies. For instance, Nightingale Open Science found that data publishers and researchers struggled with navigating a patchwork of policies from differing organizations' IRBs and data use agreements. In response, Nightingale created a Legal Toolkit[4] of a templated IRB and a data use agreement that is simpler than most others currently are to help future data publishers navigate common challenges and avoid costly mistakes. Additionally, data sharing efforts from All of Us [285], UK BioBank [276], MIMIC [132], US Centers for Medicare and Medicaid Services [62], and the US Department of Veterans' Affairs [197] have collectively shown trends in data sharing. One bright spot to further encourage is that most efforts have been moving towards cloud-based data sharing, which allows for increased security against data breaches as well as the option for equity in computing resources for researchers.

Working with the AIPF has given useful insight into a key role that academics can play, both by convening workshops of experts to distill best practices and also by sharing those recommendations with policymakers. The AIPF has built upon earlier efforts of the MIT AI Policy Congress, which brought technical experts and policymakers (including members of Congress, Presidential nominees, and civil servants) together to discuss emerging challenges that result from AI.

[4] https://www.nightingalescience.org/legal-toolkit

7.2.2 Building Norms: Privacy Watchdog

Section 7.1.3 discusses how laws often lag behind technology, sometimes by decades. Further, even when there are policies on the books, companies may not follow them [18]. This might be because of a calculated decision based on the (un-)likelihood of being caught or it might be because of norms that have developed (e.g., developers might view poor data hygeine as analogous to speeding 5 miles per hour above the speed limit).

In any matter, the legal model of accountability (e.g., law enforcement going after bad actors who violate the law), is a reactive process which occurs after tangible harm is done. To address this issue with an upstream solution (i.e., before harm occurs), Quentin Palfrey founded the International Digital Accountability Council[5] (IDAC) in 2020. During a June 2020 panel discussion with both Palfrey and Massachusetts Deputy Attorney General Sara Cable, Cable articulated that "there is an absence of real rules of the road" for mobile app developers. In response, Palfrey discussed how traditional law enforcement is only a subset of accountability tools: "I think one of the important things as you develop an ecosystem for accountability in the mobile app space is to combine the robust measures [that] the DPA or the FTC or the State AGs can do with the more nimble nonprofit actors that are still aggressive [and] still have teeth ... and really working to make sure that best practices are followed but that they have an orientation towards early intervention."

IDAC is a privacy watchdog for mobile apps, both calling out bad behavior and offering education in best practices for developers. IDAC partners with Good Research to instrument Android phones with special hardware to conduct static and dynamic tests on apps. Its investigations can reveal:

- **static**: Phone permissions requested by app.
- **static**: Software development kits present in the app.
- **dynamic**: What user data is being transmitted to and received from 3rd parties.
- **dynamic**: Who is the data going to?

In addition to technical analysis, IDAC also employs lawyers and fellows to con-

[5] Disclosure: I interned for IDAC from 2020–2021

duct policy analysis. This can take the form of grading dozens of apps' privacy policies as good or bad, reading platform (e.g., Google Play store) policies to identify violations at the platform-level, and reading FTC and State AG case law to determine whether legal action is appropriate.

When IDAC identifies a problem, it uses a "sliding scale of engagement" protocol for determining what level of intervention is appropriate.[6]

For small problems (e.g., a healthcare app using a template privacy policy which doesn't reflect the usage or purpose of how data will be used), IDAC will first try a 'polite shoulder tap' to try to resolve the issue. Many developers respond positively to the non-hostile outreach, and if they resolve the issue, then an anonymized description of the incident is reported in IDAC's "Dogs That Don't Bark" series [125] without hurting that developer's reputation.

If the developer does not fix the problem, they get publicly called out, and the pressure from users often leads to policy change. In June 2020, IDAC looked at the privacy and security practices over 100 COVID apps and called out particularly bad behavior [218]. After the report was released, many apps which we publicly flagged addressed issues by: adding privacy policies, improving privacy policies, and stopped collecting persistent identifiers. Two particularly bad apps were removed altogether.

When public pressure is not enough to convince the developers to address the problem, the next step is to identify if there are any platform (e.g., Google Play, Apple App Store) policies that they are in violation of. Private platforms can move much faster than law enforcement and may have easier-to-meet standards for the less intrusive intervention of suspending or removing an offending app. For instance, in October 2020, we flagged three apps to Google which violated the Google Play Store's developer policies that put guard rails on shadow profiling [126]. As a result, these apps were removed from the Play Store until they stopped collecting multiple identifiers that allowed them to link patient profiles across apps.

Finally, if IDAC identifies egregious misconduct by an app, they will notify law enforcement. We did this in August 2020, where we found a fertility app was collecting

[6] https://digitalwatchdog.org/policy-and-procedures-summary

personally identifying user data and sending it to China without notifying its users. As a result of this failure to disclose overseas transfer of data, we wrote a letter to both the FTC and the relevant state Attorney General asking them to investigate the company [217]. In response, Senators Warren, Klobuchar, and Moore Capito also urged the FTC to investigate the company's practices [243].

In the future, there will need to be more efforts like AIPF, IDAC, media/documentaries, and more in order to develop policies and norms around AI Ethics that are informed by lessons learned from case studies.

Appendix A

Radiology Report Model Implementation Details

A.1 Our Model Implementation

We briefly describe the details of our implementation in this section.

Encoder The image encoder CNN takes an input image of size $256 \times 256 \times 3$. The last layer before global pooling in a DenseNet-121 are extracted, which has a dimension of $8 \times 8 \times 1024$, and thus $K = 64$. Densenet-121 [124] has been shown to be state-of-the-art in the context of classification for clinical images. The image features are then projected to $d = 256$ dimensions with a dropout of $p = 0.5$.

Since typically in the X-ray image acquisition we are provided with the view position indicating the posture of the patient related to the machine, we conveniently pass this into the model as well. Indicated by a one-hot vector, the view position embedding is concatenated with the image embedding to form an input to the later decoders.

Decoder As previously mentioned, the input image embedding to the LSTM has a dimension of 256, and it is the same for word embeddings and hidden layer sizes. The word embedding matrix is pretrained with Gensim [238] in an unsupervised manner.

Training Details We implement our model on PyTorch [226] and train on 4 GeForce GTX TITAN X GPUs. All models are first trained with cross-entropy loss with the Adam [147] optimizer using an initial learning rate of 10^{-3} and a batch size of 64 for 64 epochs. Other than the weights stated above, the models are initialized randomly. Learning rates are annealed by 0.5 every 16 epochs and we increase the probability of feeding back a sample from the posterior $\mathbf{p}$ by 0.05 every 16 epochs. After this bootstrapping stage, we start training with REINFORCE for another 64 epochs. The initial learning rate for the second stage is 10^{-5} and is annealed on the same schedule.

Indicated by [240], we adopt CIDEr-D [290] metric as the reward module used in r_{NLG}. For the baseline for CCR, we choose a EMA momentum $\gamma = 0.95$. A weighting factor $\lambda = 10$ has been chosen to balance the scales of the rewards for our full model.

A.2 TieNet Re-implementation

Since the implementation for TieNet [300] is not released, we re-implement it with the descriptions provided by the original authors. The re-implementation details are described in this section.

Overview TieNet stands for *Text-Image Embedding Network*. It consists of three main components: image encoder, sentence decoder with *Attention Network*, and *Joint Learning Network*. It computes a global attention encoded text embedding using hidden states from a sentence decoder and saliency weighted global average pooling using attention maps from the attention network. The two global representations are combined as an input to the joint learning network. Finally, it outputs the multi-label classification of thoracic diseases. The end products are automatic report generation for medical images and classification of thoracic diseases.

Encoder An image of size $256 \times 256 \times 3$ is taken by the image encoder CNN as an input. The last two layers of ResNet-101 [115] are removed since we are not classifying the image. The final encoding produced has a size of $14 \times 14 \times 2048$. We also fine-tune convolutional blocks `conv2` through `conv4` of our image encoder during training time.

Decoder We also include the view position information by concatenating the view position embedding with the image embedding to form input. The view position embedding is indicated by a one-hot vector. At each decoding step, the encoded image and the previous hidden state with a dropout of $p = 0.5$ is used to generate weights for each pixel in the attention network. The previously generated word and the output from the attention network are fed to the LSTM decoder to generate the next word.

Joint Learning Network TieNet proposed an additional component to automatically classify and report thoracic diseases. The joint learning network takes hidden states and attention maps from the decoder and computes global representations for report and images, then combines the result as the input to a fully connected layer to output disease labels.

In the original paper, r indicates the number of attention heads, which we set as $r = 5$; s is the hidden size for attention generation, which we set as $s = 2000$. One key difference from the original work is that we are classifying the joint embeddings into CheXpert [127] annotated labels, and hence we have the class count $M = 14$. The disease classification cross-entropy loss L_C and the teacher-forcing report generation loss L_R are combined as $L_{\text{overall}} = \alpha L_C + (1 - \alpha) L_R$, in which L_{overall} is the loss for which the network optimizes. However, the value α was not disclosed in the original work and we use $\alpha = 0.85$.

Training We implement TieNet on PyTorch [226] and train on 4 GeForce GTX TITAN X GPUs. The decoder is trained with cross-entropy loss with the Adam [147] optimizer using an initial learning rate of 10^{-3} and a mini-batch size of 32 for 64 epochs. Learning rate for the decoder is decayed by a factor of 0.2 if there is no improvement of BLEU [222] score on the development set in 8 consecutive epochs. The joint learning network is trained with sigmoid binary cross-entropy loss with the

Adam [147] optimizer using a constant learning rate of 10^{-3}.

Result Since we are not able to access the original implementation of TieNet and we additionally inject view position information to the model, we might have small variations in result between the original paper and our re-implementation. We only compare the report generation part of TieNet to our model.

Bibliography

[1] Hal Abelson et al. "The Risks of Key Recovery, Key Escrow, and Trusted Third-Party Encryption". In: *World Wide Web J.* 2.3 (June 1997), pp. 241–257. ISSN: 1085-2301.

[2] Kendra Albert and Maggie Delano. "This Whole Thing Smacks of Gender: Algorithmic Exclusion in Bioimpedance-Based Body Composition Analysis". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 342–352. ISBN: 9781450383097. DOI: 10.1145/3442188.3445898. URL: https://doi.org/10.1145/3442188.3445898.

[3] Omar Alfarghaly et al. "Automated radiology report generation using conditioned transformers". In: *Informatics in Medicine Unlocked* 24 (2021), p. 100557. ISSN: 2352-9148. DOI: https://doi.org/10.1016/j.imu.2021.100557. URL: https://www.sciencedirect.com/science/article/pii/S2352914821000472.

[4] Emily Alsentzer et al. "Publicly Available Clinical BERT Embeddings". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 72–78. DOI: 10.18653/v1/W19-1909. URL: https://www.aclweb.org/anthology/W19-1909.

[5] AMA. *AMA history*. 2021. URL: https://www.ama-assn.org/about/ama-history/ama-history.

[6] Karim Amer and Jehane Noujaim. *The Great Hack*. 2019.

[7] Nate Anderson. *"Anonymized" data really isn't—and here's why not*. Sept. 8, 2009. URL: https://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/.

[8] Peter Anderson et al. "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Apr. 2018, pp. 6077–6086. URL: https://www.microsoft.com/en-us/research/publication/bottom-top-attention-image-captioning-visual-question-answering/.

[9] Peter Anderson et al. "SPICE: Semantic Propositional Image Caption Evaluation". In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, 2016, pp. 382–398. ISBN: 978-3-319-46454-1.

[10] Julia Angwin et al. *Machine Bias*. May 23, 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[11] Julia Angwin et al. *Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks*. 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[12] National Education Association. *Bargaining for the Common Good*. 2022. URL: https://www.nea.org/your-rights-workplace/bargaining-educator-voice/bargaining-common-good.

[13] Anand Avati et al. "Improving palliative care with deep learning". In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2017, pp. 311–316. DOI: 10.1109/BIBM.2017.8217669.

[14] Zaheer Babar, Twan van Laarhoven, and Elena Marchiori. "Encoder-decoder models for chest X-ray report generation perform no better than unconditioned baselines". In: *PLOS ONE* 16.11 (Nov. 2021), pp. 1–20. DOI: 10.1371/journal.pone.0259639. URL: https://doi.org/10.1371/journal.pone.0259639.

[15] Zaheer Babar et al. “Evaluating diagnostic content of AI-generated radiology reports of chest X-rays”. In: *Artificial Intelligence in Medicine* 116 (2021), p. 102075. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2021.102075. URL: https://www.sciencedirect.com/science/article/pii/S0933365721000683.

[16] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. English (US). In: *International Conference on Learning Representations*. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015. Jan. 2015.

[17] Dzmitry Bahdanau et al. “An actor-critic algorithm for sequence prediction”. In: *arXiv preprint arXiv:1607.07086* (2016).

[18] Kenneth A. Bamberger and Deirdre K. Mulligan. *Privacy on the Ground: Driving Corporate Behavior in the United States and Europe*. MIT Press, 2015.

[19] Satanjeev Banerjee and Alon Lavie. “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments”. In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005, pp. 65–72.

[20] Sarah Bastawrous and Benjamin Carney. “Improving Patient Safety: Avoiding Unread Imaging Exams in the National VA Enterprise Electronic Health Record”. In: *Journal of Digital Imaging* 30.3 (June 2017), pp. 309–313. ISSN: 1618-727X. DOI: 10.1007/s10278-016-9937-2. URL: https://doi.org/10.1007/s10278-016-9937-2.

[21] Haydn Belfield. “Activism by the AI Community: Analysing Recent Achievements and Future Prospects”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 15–21. ISBN: 9781450371100. URL: https://doi.org/10.1145/3375627.3375814.

[22] Emily M. Bender et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922.

[23] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. "The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database". In: *npj Digital Medicine* 3.1 (Sept. 2020), p. 118. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00324-0. URL: https://doi.org/10.1038/s41746-020-00324-0.

[24] Ruha Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code 1st Edition*. Polity, 2019. ISBN: 978-1509526406.

[25] Sam Biddle. *IBM Employees Launch Petition Protesting Cooperation With Donald Trump*. Dec. 19, 2016. URL: https://theintercept.com/2016/12/19/ibm-employees-launch-petition-protesting-cooperation-with-donald-trump.

[26] Rena Bivens. *The Gender Binary Will Not be Deprogrammed: Ten Years of Coding Gender on Facebook*. 2015. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2431443.

[27] Eric Blanc. *Red State Revolt: The Teachers' Strike Wave and Working-Class Politics*. Verso, 2019. ISBN: 978-1788735742.

[28] Bloomberg. *Google Engineers Refused to Build Security Tool to Win Military Contracts*. June 2018. URL: https://www.bloomberg.com/news/articles/2018-06-21/google-engineers-refused-to-build-security-tool-to-win-military-contracts.

[29] W. Boag et al. "Modeling Mistrust in End-of-Life Care". In: *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2018) workshop*. 2018.

[30] W. Boag et al. “Racial Disparities and Mistrust in End-of-Life Care”. In: *Proceedings of the 3rd Machine Learning for Healthcare Conference*. Vol. 85. Proceedings of Machine Learning Research. Palo Alto, California: PMLR, Aug. 2018, pp. 587–602. URL: `http://proceedings.mlr.press/v85/boag18a.html`.

[31] William Boag et al. *Healthcare Workshop Recommendations*. 2022.

[32] William Boag et al. “MUTT: Metric Unit TesTing for Language Generation Tasks”. In: *ACL*. Berlin, Germany, Aug. 2016.

[33] William Boag et al. “Tech Worker Organizing for Power and Accountability”. In: *Fairness, Accountability and Transparency*. 2022.

[34] William Boag; et al. “A Pilot Study in Surveying Clinical Judgments to Evaluate Radiology Report Generation”. In: *Fairness, Accountability and Transparency*. 2021.

[35] William Boag; et al. “Baselines for Chest X-Ray Report Generation”. In: *Machine Learning for Health workshop at NeurIPS*. 2019.

[36] Philip M. Boffey. *Software Seen As Obstacle In Developing 'Star Wars'*. Sept. 16, 1986. URL: `https://www.nytimes.com/1986/09/16/science/software-seen-as-obstacle-in-developing-star-wars.html`.

[37] Cory Booker. *Booker, Wyden, Clarke Introduce Bill Requiring Companies To Target Bias In Corporate Algorithms*. 2019. URL: `https://www.booker.senate.gov/news/press/booker-wyden-clarke-introduce-bill-requiring-companies-to-target-bias-in-corporate-algorithms`.

[38] Julie Bort. *Inside Uber before its self-driving car killed a pedestrian: Sources describe infighting, 'perverse' incentives, and questionable decisions*. 2018. URL: `https://markets.businessinsider.com/news/stocks/uber-insiders-describe-reckless-decisions-self-driving-car-unit-2018-10`.

[39] danah boyd. *Balancing Data Utility and Confidentiality in the 2020 US Census*. 2020. URL: https://datasociety.net/wp-content/uploads/2019/12/Differential-Privacy-04_27_20.pdf.

[40] Adrian P. Brady. "Error and discrepancy in radiology: inevitable or avoidable?" eng. In: *Insights into imaging* 8.1 (Feb. 2017). 27928712[pmid], pp. 171–182. ISSN: 1869-4101. DOI: 10.1007/s13244-016-0534-1. URL: https://pubmed.ncbi.nlm.nih.gov/27928712.

[41] Steven Brill. *Class Warfare: Inside the Fight to Fix America's Schools*. Simon and Schuster, 2012.

[42] Meredith Broussard. *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press, 2018. ISBN: 978-0262537018.

[43] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. New York, NY, USA: PMLR, Feb. 2018, pp. 77–91. URL: http://proceedings.mlr.press/v81/buolamwini18a.html.

[44] Katelyn Burns. *The racist history of Trump's "When the looting starts, the shooting starts" tweet*. May 29, 2020. URL: https://www.vox.com/identities/2020/5/29/21274754/racist-history-trump-when-the-looting-starts-the-shooting-starts.

[45] Aurelia Bustos et al. "Padchest: A large chest x-ray image dataset with multi-label annotated reports". In: *arXiv preprint arXiv:1901.07441* (2019).

[46] Massimo Caccia et al. "Language gans falling short". In: *arXiv preprint arXiv:1811.02549* (2018).

[47] Matt Cagle and Nicole Ozer. *Amazon Teams Up With Government to Deploy Dangerous New Facial Recognition Technology*. May 22, 2018. URL: https://

www.aclu.org/blog/privacy-technology/surveillance-technologies/amazon-teams-government-deploy-dangerous-new.

[48] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. "Re-evaluation the role of bleu in machine translation research". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. 2006.

[49] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. "Re-evaluation the role of bleu in machine translation research". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. 2006.

[50] Nicholas Carlini et al. "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks". In: *ArXiv e-prints* 1802.08232 (2018). URL: https://arxiv.org/abs/1802.08232.

[51] CDC. *History of Smallpox*. 2021. URL: https://www.cdc.gov/smallpox/history/history.html.

[52] CDC. *The Tuskegee Timeline*. 2021. URL: https://www.cdc.gov/tuskegee/timeline.htm.

[53] Jeanne Sternlicht Chall and Edgar Dale. *Readability revisited: the new Dale-Chall readability formula*. Brookline Books, 1995.

[54] Rosalie Chan. *Read the internal letter sent by a group of Amazon employees asking the company to take a stand against ICE*. July 11, 2019. URL: https://www.businessinsider.com/amazon-employees-letter-protest-palantir-ice-camps-2019-7.

[55] Zhengping Che et al. "Recurrent neural networks for multivariate time series with missing values". In: *Scientific reports* 8.1 (2018), p. 6085.

[56] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

[57] Zhihong Chen et al. “Generating Radiology Reports via Memory-driven Transformer”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Nov. 2020.

[58] Marc Cheong, Kobi Leins, and Simon Coghlan. “Computer Science Communities: Who is Speaking, and Who is Listening to the Women? Using an Ethics of Care to Promote Diverse Voices”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 106–115. ISBN: 9781450383097. DOI: 10.1145/3442188.3445874. URL: https://doi.org/10.1145/3442188.3445874.

[59] Edward Choi et al. “Generating Multi-label Discrete Patient Records using Generative Adversarial Networks”. In: *Proceedings of the 2nd Machine Learning for Healthcare Conference*. Vol. 68. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 286–305. URL: https://proceedings.mlr.press/v68/choi17a.html.

[60] Alexandra Chouldechova et al. “A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions”. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, Feb. 2018, pp. 134–148. URL: https://proceedings.mlr.press/v81/chouldechova18a.html.

[61] "Amazon Employees for Climate Justice". *Open letter to Jeff Bezos and the Amazon Board of Directors*. Apr. 10, 2019. URL: https://amazonemployees4climatejustice.medium.com/public-letter-to-jeff-bezos-and-the-amazon-board-of-directors-82a8405f5e38.

[62] CMS. Accessed: 2022-05-12. 2004. URL: https://www.cms.gov/Research-Statistics-Data-and-Systems/Research-Statistics-Data-and-Systems.

[63] Aloni Cohen and Kobbi Nissim. “Towards formalizing the GDPR’s notion of singling out”. In: *Proceedings of the National Academy of Sciences* 117.15

(2020), pp. 8344–8352. ISSN: 0027-8424. DOI: 10.1073/pnas.1914598117. eprint: https://www.pnas.org/content/117/15/8344.full.pdf. URL: https://www.pnas.org/content/117/15/8344.

[64] C. D'Ignazio and L.F. Klein. *Data Feminism*. Strong Ideas. MIT Press, 2020. ISBN: 9780262044004. URL: https://books.google.com/books?id=x5nSDwAAQBAJ.

[65] Jeffrey R. Darst et al. "Deciding without Data". In: *Congenital Heart Disease* 5 (July 2010), pp. 339–342. DOI: 10.1111/j.1747-0803.2010.00433.x.

[66] Anupam Datta, Divya Sharma, and Arunesh Sinha. "Provable De-anonymization of Large Datasets with Sparse Dimensions". In: Apr. 2012. ISBN: 978-3-642-28640-7. DOI: 10.1007/978-3-642-28641-4_13.

[67] Dina Demner-Fushman et al. "Preparing a collection of radiology examinations for distribution and retrieval". In: *Journal of the American Medical Informatics Association* 23.2 (2016), pp. 304–310.

[68] Jacob Devlin et al. "Exploring nearest neighbor approaches for image captioning". In: *arXiv preprint arXiv:1505.04467* (2015).

[69] Megan Rose Dickey. *The weaponization of employee resource groups*. Sept. 21, 2021. URL: https://www.protocol.com/workplace/employee-resource-group-weaponization%5C#toggle-gdpr.

[70] Trafton Drew, Melissa L.-H. Võ, and Jeremy M. Wolfe. "The Invisible Gorilla Strikes Again: Sustained Inattentional Blindness in Expert Observers". In: *Psychological Science* 24.9 (Sept. 2013), pp. 1848–1853. ISSN: 0956-7976. DOI: 10.1177/0956797613479386. URL: https://doi.org/10.1177/0956797613479386.

[71] *Differential privacy*. Vol. 2006. ICALP, 2006, pp. 1–12. URL: https://link.springer.com/chapter/10.1007/11787006_1.

[72] Elizabeth Dwoskin, Nitasha Tiku, and Craig Timberg. *Facebook's race-blind practices around hate speech came at the expense of Black users, new docu-*

ments show. 2021. URL: https://www.washingtonpost.com/technology/2021/11/21/facebook-algorithm-biased-race.

[73] Elizabeth Dwoskin, Nitasha Tiku, and Craig Timberg. *Facebook's race-blind practices around hate speech came at the expense of Black users, new documents show.* Nov. 21, 2021. URL: https://www.washingtonpost.com/technology/2021/11/21/facebook-algorithm-biased-race.

[74] Alistair E W Johnson, Andrew Kramer, and Gari D Clifford. "A New Severity of Illness Scale Using a Subset of Acute Physiology, Age, and Chronic Health Evaluation Data Elements Shows Comparable Predictive Accuracy." In: 41 (May 2013).

[75] David M. Eddy. "Practice Policies—Guidelines for Methods". In: *JAMA* 263 (1990), pp. 1839–1841.

[76] Khaled El Emam et al. "A Systematic Review of Re-Identification Attacks on Health Data". In: *PLOS ONE* 6.12 (Dec. 2011), pp. 1–12. DOI: 10.1371/journal.pone.0028071. URL: https://doi.org/10.1371/journal.pone.0028071.

[77] Anne Elixhauser et al. "Comorbidity measures for use with administrative data." In: *Medical care* 36 1 (1998), pp. 8–27.

[78] Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* Picador, 2018. ISBN: 978-1250215789.

[79] Tasha Eurich. *Insight: The Surprising Truth About How Others See Us, How We See Ourselves, and Why the Answers Matter More Than We Think.* Currency, 2018. ISBN: 978-0525573944.

[80] Lee Fang. *Leaked Emails Show Google Expected Lucrative Military Drone to Work to Grow Exponentially.* May 31, 2018. URL: https://theintercept.com/2018/05/31/google-leaked-emails-drone-ai-pentagon-lucrative.

[81] Megan Farokhmanesh. *Google warns its employees that Pride protests are against the company's code of conduct.* June 24, 2019. URL: https://www.theverge.com/2019/6/24/18716204/google-employees-pride-protest-code-of-conduct-violation.

[82] FDA. *Artificial Intelligence/Machine Learning (AI/ML)-Based.:Jf/<X Software as a Medical Device (SaMD) Action Plan.* 2021. URL: https://www.fda.gov/media/145022/download.

[83] Todd Feathers. *Google's New Dermatology App Wasn't Designed for People With Darker Skin.* May 20, 2021. URL: https://www.vice.com/en/article/m7evmy/googles-new-dermatology-app-wasnt-designed-for-people-with-darker-skin.

[84] William Fedus, Ian Goodfellow, and Andrew M Dai. "Maskgan: Better text generation via filling in the_". In: *arXiv preprint arXiv:1801.07736* (2018).

[85] Rudolph Flesch. *A new readability yardstick.* US, 1948. DOI: 10.1037/h0057532.

[86] Sam Fletcher and Md Zahidul Islam. "Differentially private random decision forests using smooth sensitivity". In: *Expert Systems with Applications* 78 (2017), pp. 16–31. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2017.01.034. URL: https://www.sciencedirect.com/science/article/pii/S0957417417300428.

[87] Sheera Frenkel and Cecilia Kang. *An Ugly Truth: Inside Facebook's Battle for Domination.* Harper, 2021.

[88] Sorelle A. Friedler and Christo Wilson. "Preface". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency.* Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, Feb. 2018, pp. 1–2. URL: https://proceedings.mlr.press/v81/friedler18a.html.

[89] Sorelle A. Friedler et al. “A Comparative Study of Fairness-Enhancing Interventions in Machine Learning”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 329–338. ISBN: 9781450361255. DOI: 10.1145/3287560.3287589. URL: https://doi.org/10.1145/3287560.3287589.

[90] Joseph Futoma et al. “An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection”. In: *Proceedings of the 2nd Machine Learning for Healthcare Conference*. Ed. by Finale Doshi-Velez et al. Vol. 68. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 243–254. URL: https://proceedings.mlr.press/v68/futoma17a.html.

[91] William Gale et al. “Producing radiologist-quality reports for interpretable artificial intelligence”. In: *arXiv preprint arXiv:1806.00340* (2018).

[92] Julia Galef. *The Scout Mindset: Why Some People See Things Clearly and Others Don’t*. Portfolio, 2021. ISBN: 978-0735217553.

[93] Alba Garcı a Seco de Herrera et al. “Overview of the ImageCLEF 2018 caption prediction tasks”. In: *Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum (CLEF 2018), Avignon, France, September 10-14, 2018.* Vol. 2125. CEUR Workshop Proceedings. 2018.

[94] Simson Garfinkel, John M. Abowd, and Christian Martindale. “Understanding Database Reconstruction Attacks on Public Data”. In: *Commun. ACM* 62.3 (Feb. 2019), pp. 46–53. ISSN: 0001-0782. DOI: 10.1145/3287287. URL: https://doi.org/10.1145/3287287.

[95] Joanne Mills Garrett et al. “Life-sustaining treatments during terminal illness - Who wants what?” In: *Journal of General Internal Medicine* 8.7 (July 1993), pp. 361–368. ISSN: 0884-8734. DOI: 10.1007/BF02600073.

[96] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. *The Perpetual Line-Up: Unregulated Police Face Recognition in America*. 2016. URL: https://www.perpetuallineup.org/.

[97] Atul Gawande. *Being Mortal: Medicine and What Matters in the End*. Picador, 2014.

[98] Timnit Gebru et al. *Datasheets for Datasets*. Mar. 2018. URL: https://www.microsoft.com/en-us/research/publication/datasheets-for-datasets.

[99] Shirin Ghaffary. *GitHub is the latest tech company to face controversy over its contracts with ICE*. Oct. 9, 2019. URL: https://www.vox.com/recode/2019/10/9/20906605/github-ice-contract-immigration-ice-dan-friedman.

[100] Shirin Ghaffary. *Google says it's committed to ethical AI research. Its ethical AI team isn't so sure*. June 2, 2021. URL: https://www.vox.com/recode/22465301/google-ethical-ai-timnit-gebru-research-alex-hanna-jeff-dean-marian-croak.

[101] Amy Goodman and Denis Moynihan. *LA teachers strike a blow against school privatization*. Jan. 25, 2019. URL: https://captimes.com/opinion/column/amy-goodman-and-denis-moynihan-la-teachers-strike-a-blow-against-school-privatization/article_f761e5a4-1b60-5430-8186-1de3fbfe674f.html.

[102] Kevin Granville. *Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens*. 2018. URL: https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html.

[103] Alex Graves. "Generating sequences with recurrent neural networks". In: *arXiv preprint arXiv:1308.0850* (2013).

[104] Andy Greenberg. *How One of Apple's Key Privacy Safeguards Falls Short*. Sept. 15, 2017. URL: https://www.wired.com/story/apple-differential-privacy-shortcomings.

[105] Jay Greene. *Amazon fires two tech workers who criticized the company's warehouse workplace conditions*. Apr. 14, 2020. URL: https://www.washingtonpost.com/technology/2020/04/13/amazon-workers-fired.

[106] Jay Greene. *Amazon settles unfair labor claims with two fired tech workers*. Sept. 29, 2021. URL: https://www.washingtonpost.com/technology/2021/09/29/amazon-settlement-fired-workers.

[107] Robert Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1968. ISBN: 0070252068.

[108] Hans Haferkamp and Neil J Smelser. *Social Change and Modernity*. Berkeley, California: University of California Press, 1992.

[109] Amresh Hanchate et al. "Racial and ethnic differences in end-of-life costs: Why do minorities cost more than whites?" In: *Archives of Internal Medicine* 169.5 (2009), pp. 493–501.

[110] Karen Hao. *Should a self-driving car kill the baby or the grandma? Depends on where you're from*. 2018. URL: https://www.technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem.

[111] Karen Hao. *The two-year fight to stop Amazon from selling face recognition to the police*. 2020. URL: https://www.technologyreview.com/2020/06/12/1003482/amazon-stopped-selling-police-face-recognition-fight.

[112] Albert Haque et al. "Towards Vision-Based Smart Hospitals: A System for Tracking and Monitoring Hand Hygiene Compliance". In: *Proceedings of the 2nd Machine Learning for Healthcare Conference*. Ed. by Finale Doshi-Velez et al. Vol. 68. Proceedings of Machine Learning Research. PMLR, Sept. 2017, pp. 75–87. URL: https://proceedings.mlr.press/v68/haque17a.html.

[113] Sandra Harding. ""Strong Objectivity": A Response to the New Objectivity Question". In: *Synthese* 104.3 (1995), pp. 331–349. ISSN: 00397857, 15730964. URL: http://www.jstor.org/stable/20117437.

[114] Joshua M. Hauser et al. "Minority populations and advance directives: Insights from a focus group methodology". In: *Cambridge Quarterly of Healthcare Ethics* 6.1 (1997), pp. 58–71. ISSN: 0963-1801. DOI: 10.1017/S0963180100007611.

[115] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[116] Dan Heath. *Upstream: The Quest to Solve Problems Before They Happen*. Avid Reader Press / Simon & Schuster, 2020. ISBN: 978-1982134723.

[117] Dan Heath and Chip Heath. *Switch: How to Change Things When Change Is Hard*. Crown Business, 2010. ISBN: 978-0385528757.

[118] Alex Hern. *Fitness tracking app Strava gives away location of secret US army bases*. Jan. 28, 2018. URL: https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases.

[119] Vaughn Hillyard. *Donald Trump’s Plan for a Muslim Database Draws Comparison to Nazi Germany*. Nov. 19, 2015. URL: ttps://www.nbcnews.com/politics/2016-election/trump-says-he-would-certainly-implement-muslim-database-n466716.

[120] Naoise Holohan et al. “Diffprivlib: the IBM differential privacy library”. In: *ArXiv e-prints* 1907.02444 [cs.CR] (July 2019).

[121] Daibing Hou et al. “Automatic Report Generation for Chest X-Ray Images via Adversarial Reinforcement Learning”. In: *IEEE Access* 9 (2021), pp. 21236–21250. DOI: 10.1109/ACCESS.2021.3056175.

[122] Tzu-Ming Harry Hsu et al. “Unsupervised Multimodal Representation Learning across Medical Images and Reports”. In: *arXiv preprint arXiv:1811.08615* (2018).

[123] Xin Huang et al. “Multi-Attention and Incorporating Background Information Model for Chest X-Ray Image Report Generation”. In: *IEEE Access* 7 (2019), pp. 154808–154817. DOI: 10.1109/ACCESS.2019.2947134.

[124] Forrest Iandola et al. “Densenet: Implementing efficient convnet descriptor pyramids”. In: *arXiv preprint arXiv:1404.1869* (2014).

[125] IDAC. *Dogs That Don't Bark: Resolving a COVID-19 Symptom Checker App's Privacy Policy Issues.* June 18, 2020. URL: https://digitalwatchdog.org/dogs-that-dont-bark-resolving-a-covid-19-symptom-checker-apps-privacy-policy-issues.

[126] IDAC. *IDAC Researchers Alert Google to Policy-Violating Third-Party Data Practices.* 2020. URL: https://digitalwatchdog.org/idac-researchers-alert-google-to-policy-violating-third-party-data-practices.

[127] Jeremy Irvin et al. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison". In: *arXiv preprint arXiv:1901.07031* (2019).

[128] Zachary Izzo et al. "Approximate Data Deletion from Machine Learning Models". In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics.* Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, Apr. 2021, pp. 2008–2016. URL: https://proceedings.mlr.press/v130/izzo21a.html.

[129] Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan V.S. "Automatic Detection of Machine Generated Text: A Critical Survey". In: *Proceedings of the 28th International Conference on Computational Linguistics.* Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 2296–2309. DOI: 10.18653/v1/2020.coling-main.208. URL: https://aclanthology.org/2020.coling-main.208.

[130] Thomas Jefferson. *Correspondence with John Vaughn.* 1801.

[131] Baoyu Jing, Pengtao Xie, and Eric Xing. "On the Automatic Generation of Medical Imaging Reports". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 2577–2586. DOI: 10.18653/v1/P18-1240. URL: https://aclanthology.org/P18-1240.

[132] Alistair E W Johnson et al. “MIMIC-III, a freely accessible critical care database”. In: *Scientific data* 3 (2016).

[133] Alistair EW Johnson et al. “MIMIC-CXR: A large publicly available database of labeled chest radiographs”. In: *arXiv preprint arXiv:1901.07042* 1.2 (2019).

[134] Khari Johnson. *AI ethics research conference suspends Google sponsorship*. Mar. 2, 2021. URL: https://venturebeat.com/2021/03/02/ai-ethics-research-conference-suspends-google-sponsorship.

[135] Kimberly S. Johnson. “Racial and ethnic disparities in palliative care”. eng. In: *Journal of palliative medicine* 16.11 (Nov. 2013), pp. 1329–1334. ISSN: 1557-7740. DOI: 10.1089/jpm.2013.9468. URL: https://pubmed.ncbi.nlm.nih.gov/24073685.

[136] Alan E Jones, Stephen Trzeciak, and Jeffrey A Kline. “The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation”. In: *Critical care medicine* 37.5 (May 2009), pp. 1649–1654. ISSN: 0090-3493. DOI: 10.1097/ccm.0b013e31819def97. URL: https://europepmc.org/articles/PMC2703722.

[137] Kristopher N. Jones et al. “PEIR Digital Library: Online Resources and Authoring System”. eng. In: *Proceedings of the AMIA Symposium* (2001). PMC2243692[pmcid], pp. 1075–1075. ISSN: 1531-605X. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243692/.

[138] Peter Kafka. *Why the best referees for Twitter and Facebook may be the people who work there.* Jan. 14, 2021. URL: https://www.vox.com/recode/22230903/trump-riot-twitter-facebook-employees-qanon-kevin-roose-ben-collins-recode-media-podcast.

[139] Richard D. Kahlenberg. *Bipartisan, But Unfounded: The Assault on Teachers’ Unions.* 2011. URL: https://www.aft.org/sites/default/files/periodicals/Kahlenberg_0.pdf.

[140] Hassan Kane et al. "NUBIA: NeUral Based Interchangeability Assessor for Text Generation". In: *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*. Online (Dublin, Ireland): Association for Computational Linguistics, Dec. 2020, pp. 28–37. URL: https://www.aclweb.org/anthology/2020.evalnlgeval-1.4.

[141] Hassan Kané et al. "Towards Neural Similarity Evaluator". In: *Workshop on Document Intelligence at NeurIPS 2019*. 2019.

[142] Cecilia Kang and Mike Isaac. *Defiant Zuckerberg Says Facebook Won't Police Political Speech*. Oct. 17, 2019. URL: https://www.nytimes.com/2019/10/17/business/zuckerberg-facebook-free-speech.html.

[143] Shalini Kantayya. *Coded Bias*. 2021.

[144] Makena Kelly. *Congress is way behind on algorithmic misinformation*. 2021. URL: https://www.theverge.com/2021/4/27/22406054/facebook-twitter-google-youtube-algorithm-transparency-regulation-misinformation-disinformation.

[145] Mert Kilickaya et al. "Re-evaluating Automatic Metrics for Image Captioning". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 199–209. URL: https://www.aclweb.org/anthology/E17-1019.

[146] Kunho Kim et al. "Differentially Private n-gram Extraction". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 5102–5111. URL: https://proceedings.neurips.cc/paper/2021/file/28ce9bc954876829eeb56ff46da8e1ab-Paper.pdf.

[147] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[148] Ezra Klein. *Mark Zuckerberg on Facebook's hardest year, and what comes next.* Apr. 2018.

[149] Ezra Klein. *Why We're Polarized.* Avid Reader Press / Simon & Schuster, 2020. ISBN: 978-1476700328.

[150] Will Knight. *These Doctors Are Using AI to Screen for Breast Cancer.* 2021. URL: https://www.wired.com/story/doctors-using-ai-screen-breast-cancer.

[151] Vasiliki Kougia et al. "RTEX: A novel framework for ranking, tagging, and explanatory diagnostic captioning of radiography exams". In: *Journal of the American Medical Informatics Association* 28.8 (Apr. 2021), pp. 1651–1659. ISSN: 1527-974X. DOI: 10.1093/jamia/ocab046. eprint: https://academic.oup.com/jamia/article-pdf/28/8/1651/39502314/ocab046.pdf. URL: https://doi.org/10.1093/jamia/ocab046.

[152] P. M. Krafft et al. "An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 772–781. ISBN: 9781450383097. DOI: 10.1145/3442188.3445938. URL: https://doi.org/10.1145/3442188.3445938.

[153] Anna Kramer. *Mobilize app workers have unionized, adding momentum to CWA's tech organizing efforts.* Mar. 15, 2021. URL: https://www.protocol.com/mobilize-app-code-cwa-union.

[154] Anna Kramer. *The Glitch union just signed the first tech company collective bargaining agreement.* Mar. 2, 2021. URL: https://www.protocol.com/bulletins/glitch-union-collective-bargaining.

[155] Jonathan Krause et al. "A hierarchical approach for generating descriptive image paragraphs". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2017, pp. 317–325.

[156] Hei Law, Khurshid Ghani, and Jia Deng. “Surgeon Technical Skill Assessment using Computer Vision based Analysis”. In: *Proceedings of the 2nd Machine Learning for Healthcare Conference*. Ed. by Finale Doshi-Velez et al. Vol. 68. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 88–99. URL: https://proceedings.mlr.press/v68/law17a.html.

[157] Crystal Lee et al. “Viral Visualizations: How Coronavirus Skeptics Use Orthodox Data Practices to Promote Unorthodox Science Online”. In: *ACM Human Factors in Computing Systems (CHI)*. 2021. DOI: 10.1145/3411764.3445211. URL: http://vis.csail.mit.edu/pubs/viral-visualizations.

[158] Janet J. Lee et al. “The Influence of Race/Ethnicity and Education on Family Ratings of the Quality of Dying in the ICU”. In: *Journal of Pain and Symptom Management* 51.1 (2016), pp. 9–16. ISSN: 0885-3924. DOI: https://doi.org/10.1016/j.jpainsymman.2015.08.008. URL: http://www.sciencedirect.com/science/article/pii/S0885392415004558.

[159] Nicol Turner Lee, Jack Karsten, and Jordan Roberts. *Removing regulatory barriers to telehealth before and after COVID-19*. Tech. rep. Brookings Institute, 2020. URL: https://www.brookings.edu/wp-content/uploads/2020/05/Removing-barriers-to-telehealth-before-and-after-COVID-19_PDF.pdf.

[160] Barron H. Lerner. *Scholars Argue Over Legacy of Surgeon Who Was Lionized, Then Vilified*. 2003.

[161] Jiwei Li et al. “Deep reinforcement learning for dialogue generation”. In: *arXiv preprint arXiv:1606.01541* (2016).

[162] Yuan Li et al. “Hybrid retrieval-generation reinforced agent for medical image report generation”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 1537–1547.

[163] Chin-Yew Lin. “ROUGE: a Package for Automatic Evaluation of Summaries”. In: *Workshop on Text Summarization Branches Out*. July 2004. URL: https:

//www.microsoft.com/en-us/research/publication/rouge-a-package-for-automatic-evaluation-of-summaries/.

[164] Chin-Yew Lin. “Rouge: A package for automatic evaluation of summaries”. In: *Text Summarization Branches Out* (2004).

[165] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *ECCV*. European Conference on Computer Vision, Sept. 2014. URL: https://www.microsoft.com/en-us/research/publication/microsoft-coco-common-objects-in-context/.

[166] Geert Litjens et al. “A survey on deep learning in medical image analysis”. In: *Medical image analysis* 42 (2017), pp. 60–88.

[167] Fenglin Liu et al. “Contrastive Attention for Automatic Chest X-ray Report Generation”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 269–280. DOI: 10.18653/v1/2021.findings-acl.23. URL: https://aclanthology.org/2021.findings-acl.23.

[168] Guanxiong Liu et al. “Clinically Accurate Chest X-Ray Report Generation”. In: *Proceedings of the 4th Machine Learning for Healthcare Conference*. Vol. 106. Proceedings of Machine Learning Research. Ann Arbor, Michigan: PMLR, Aug. 2019, pp. 249–269. URL: http://proceedings.mlr.press/v106/liu19a.html.

[169] Justin Lovelace and Bobak Mortazavi. “Learning to Generate Clinically Coherent Chest X-Ray Reports”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1235–1243. DOI: 10.18653/v1/2020.findings-emnlp.110. URL: https://aclanthology.org/2020.findings-emnlp.110.

[170] Donna Lu. *Doctor ‘bias’ behind women getting worse treatment for heart attacks, Australian study finds.* 2021. URL: https://www.theguardian.com/australia-news/2021/sep/20/doctor-bias-behind-women-getting-worse-treatment-for-heart-attacks-australian-study-finds.

[171] Jiasen Lu et al. “Knowing when to look: Adaptive attention via a visual sentinel for image captioning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2017, pp. 375–383.

[172] Alexandra Ma and Ben Gilbert. *Facebook understood how dangerous the Trump-linked data firm Cambridge Analytica could be much earlier than it previously said. Here’s everything that’s happened up until now.* 2019. URL: `https://www.businessinsider.com/cambridge-analytica-a-guide-to-the-trump-linked-data-firm-that-harvested-50-million-facebook-profiles-2018-3%5C#does-this-change-facebook-and-cambridge-analyticas-previous-testimonies-6`.

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

[174] William MacAskill. *Doing Good Better.* Avery, 2016. ISBN: 978-1592409662.

[175] A. Machanavajjhala et al. “L-diversity: privacy beyond k-anonymity”. In: *22nd International Conference on Data Engineering (ICDE’06).* 2006, pp. 24–24. DOI: `10.1109/ICDE.2006.1`.

[176] Michael A. Madaio et al. “Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–14. ISBN: 9781450367080. URL: `https://doi.org/10.1145/3313831.3376445`.

[177] M. Marelli et al. “A SICK cure for the evaluation of compositional distributional semantic models”. In: *Proceedings of LREC 2014.* 2014, pp. 216–223.

[178] Jane McAlevey. “No Shortcuts: The Case for Organizing”. PhD thesis. City University of New York, 2015.

[179] Matthew B.A. McDermott et al. “CheXpert++: Approximating the CheXpert Labeler for Speed, Differentiability, and Probabilistic Output”. In: *Proceedings of the 5th Machine Learning for Healthcare Conference*. Ed. by Finale Doshi-Velez et al. Vol. 126. Proceedings of Machine Learning Research. PMLR, Aug. 2020, pp. 913–927. URL: https://proceedings.mlr.press/v126/mcdermott20a.html.

[180] Kelly McGonigal. *The Science of Compassion: A Modern Approach for Cultivating Empathy, Love, and Connection*. Sounds True, 2016. ISBN: 978-1622037797.

[181] Emma Meats et al. “Evidence-based medicine teaching in UK medical schools”. In: *Medical Teacher* 31.4 (2009), pp. 332–337. DOI: 10.1080/01421590802572791.

[182] Pablo Messina et al. *A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images*. 2022. arXiv: 2010.10563 [cs.CV].

[183] Milagros Miceli et al. “Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 161–172. ISBN: 9781450383097. DOI: 10.1145/3442188.3445880. URL: https://doi.org/10.1145/3442188.3445880.

[184] Margaret Mitchell et al. “Model Cards for Model Reporting”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 220–229. ISBN: 9781450361255. DOI: 10.1145/3287560.3287596. URL: https://doi.org/10.1145/3287560.3287596.

[185] Margaret Mitchell et al. “Model Cards for Model Reporting”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 220–229. ISBN: 9781450361255. DOI: 10.1145/3287560.3287596. URL: https://doi.org/10.1145/3287560.3287596.

[186] Shira Mitchell et al. “Algorithmic Fairness: Choices, Assumptions, and Definitions”. In: *Annual Review of Statistics and Its Application* 8.1 (2021), pp. 141–163. DOI: 10.1146/annurev-statistics-042720-125902. URL: https://doi.org/10.1146/annurev-statistics-042720-125902.

[187] Brent Mittelstadt, Chris Russell, and Sandra Wachter. “Explaining Explanations in AI”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency.* FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 279–288. ISBN: 9781450361255. DOI: 10.1145/3287560.3287574. URL: https://doi.org/10.1145/3287560.3287574.

[188] Yasuhide Miura et al. “Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Online: Association for Computational Linguistics, June 2021, pp. 5288–5304. DOI: 10.18653/v1/2021.naacl-main.416. URL: https://aclanthology.org/2021.naacl-main.416.

[189] Alan Mossman. *Using Plus/Delta for Feedback and Improving Social Processes.* Sept. 12, 2019. URL: https://leanconstructionblog.com/Using-Plus-Delta-for-Feedback-and-Improving-Social-Processes.html.

[190] Sarah Muni et al. “The Influence of Race/Ethnicity and Socioeconomic Status on End-of-Life Care in the ICU”. In: *Chest* 139.5 (2011), pp. 1025–1033. ISSN: 0012-3692. DOI: https://doi.org/10.1378/chest.10-3011. URL: https://www.sciencedirect.com/science/article/pii/S0012369211602262.

[191] n/a. *Creating Diverse & Inclusive Business: The Evolution of Employee Resource Groups.* June 18, 2020. URL: https://www.realizedworth.com/2019/06/26/creating-diverse-inclusive-business-the-evolution-of-employee-resource-groups.

[192] Ivona Najdenkoska et al. “Variational Topic Inference for Chest X-Ray Report Generation”. In: *CoRR* abs/2107.07314 (2021). arXiv: 2107.07314. URL: https://arxiv.org/abs/2107.07314.

[193] Arvind Narayanan and Vitaly Shmatikov. "Robust De-anonymization of Large Sparse Datasets". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. 2008, pp. 111–125. DOI: 10.1109/SP.2008.33.

[194] Arvind Narayanan and Vitaly Shmatikov. "Robust de-anonymization of large sparse datasets: a decade later". In: May 21, 2019.

[195] Arvind Narayanan and Vitaly Shmatikov. *Robust de-anonymization of large sparse datasets: a decade later*. 2019. URL: https://www.cs.princeton.edu/~arvindn/publications/de-anonymization-retrospective.pdf.

[196] Arvind Narayanan et al. "Dark patterns". English (US). In: *Communications of the ACM* 63.9 (Sept. 2020). Publisher Copyright: © 2020 ACM., pp. 42–47. ISSN: 0001-0782. DOI: 10.1145/3397884.

[197] National Center for Veterans Analysis and Statistics. Accessed: 2022-05-12. 2022. URL: https://www.va.gov/vetdata.

[198] National Technical Information Service. *IMPORTANT NOTICE: Change in Public Death Master File Records*. 2011. URL: https://ladmf.ntis.gov/docs/import-change-dmf.pdf.

[199] Nataliya Nedzhvetskaya and JS Tan. *In Oxford Handbook on AI Governance: The Role of Workers in AI Ethics and Governance*. Aug. 2021.

[200] Bret Nestor et al. "Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks". In: *Proceedings of the 4th Machine Learning for Healthcare Conference*. Ed. by Finale Doshi-Velez et al. Vol. 106. Proceedings of Machine Learning Research. PMLR, Aug. 2019, pp. 381–405. URL: https://proceedings.mlr.press/v106/nestor19a.html.

[201] Ha Q. Nguyen et al. *VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations*. 2020. arXiv: 2012.15029 [eess.IV].

[202] Hoang Nguyen et al. "Automated Generation of Accurate & Fluent Medical X-ray Reports". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3552–3569. DOI: `10.18653/v1/2021.emnlp-main.288`. URL: `https://aclanthology.org/2021.emnlp-main.288`.

[203] Jianmo Ni et al. "Learning Visual-Semantic Embeddings for Reporting Abnormal Findings on Chest X-rays". In: *Findings of the Association for Computational Linguistics: EMNLP 2020.* Online: Association for Computational Linguistics, Nov. 2020, pp. 1954–1960. DOI: `10.18653/v1/2020.findings-emnlp.176`. URL: `https://aclanthology.org/2020.findings-emnlp.176`.

[204] Aaron Nicolson, Jason Dowling, and Bevan Koopman. *Improving Chest X-Ray Report Generation by Leveraging Warm-Starting.* 2022. arXiv: `2201.09405 [cs.CV]`.

[205] Toru Nishino et al. "Reinforcement Learning with Imbalanced Dataset for Data-to-Text Medical Report Generation". In: *Findings of the Association for Computational Linguistics: EMNLP 2020.* Online: Association for Computational Linguistics, Nov. 2020, pp. 2223–2236. DOI: `10.18653/v1/2020.findings-emnlp.202`. URL: `https://aclanthology.org/2020.findings-emnlp.202`.

[206] Safiya Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press, 2018. ISBN: 978-1479837243.

[207] Jekaterina Novikova et al. "Why We Need New Evaluation Metrics for NLG". English. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Sept. 2017, pp. 2231–2242.

[208] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown, 2016. ISBN: 978-0553418835.

[209] L. Oakden-Raynor. *The Medical AI Floodgates Open, at a Cost of $1000 per Patient*. Tech. rep. The Healthcare Blog, 2020. URL: https://thehealthcareblog.com/blog/2020/09/10/the-medical-ai-floodgates-open-at-a-cost-of-1000-per-patient.

[210] Ziad Obermeyer and Sendhil Mullainathan. "Dissecting Racial Bias in an Algorithm That Guides Health Decisions for 70 Million People". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, 2019, p. 89. ISBN: 9781450361255. DOI: 10.1145/3287560.3287593. URL: https://doi.org/10.1145/3287560.3287593.

[211] OECD. *OECD Principles for Internet Policymaking*. 2014. URL: http://www.oecd.org/sti/ieconomy/oecd-principles-for-internet-policy-making.pdf.

[212] Omer Onder et al. "Errors, discrepancies and underlying bias in radiology with case examples: a pictorial review". In: *Insights into Imaging* 12.1 (Apr. 2021), p. 51. ISSN: 1869-4101. DOI: 10.1186/s13244-021-00986-8. URL: https://doi.org/10.1186/s13244-021-00986-8.

[213] Jeff Orlowski. *The Social Dilemmna*. 2020.

[214] Alexandra Ossola. *Are Engineers Responsible for the Consequences of Their Algorithms?* Dec. 8, 2017. URL: https://futurism.com/are-engineers-responsible-for-the-consequences-of-their-algorithms.

[215] Leila Ouchchy, Allen Coin, and Veljko Dubljevi. "AI in the Headlines: The Portrayal of the Ethical Issues of Artificial Intelligence in the Media". In: *AI and Society* 35.4 (2020), pp. 927–936. DOI: 10.1007/s00146-020-00965-5.

[216] Nitish Pahwa. *How Morale at Facebook Tumbled After Trump's "Looting and Shooting" Post*. Nov. 30, 2021. URL: https://slate.com/technology/2021/11/facebook-internal-employee-survey-morale-trump-looting-shooting.html.

[217] Quentin Palfrey and Lena Ghamrawi. *Re: Premom's Deceptive Privacy Practices Places Vulnerable Users' Data at Risk*. 2020. URL: https://digitalwatchdog.org/wp-content/uploads/2020/08/IDAC-Federal-Trade-Commission-Letter.pdf.

[218] Quentin Palfrey et al. *Privacy in the Age of COVID: An IDAC Investigation of COVID-19 Apps*. July 10, 2020. URL: https://digitalwatchdog.org/wp-content/uploads/2020/07/IDAC-COVID19-Mobile-Apps-Investigation-07132020.pdf.

[219] Nicolas Papernot et al. *Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data*. 2017. DOI: 10.48550/ARXIV.1610.05755.

[220] Kishore Papineni et al. "BLEU: A Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://doi.org/10.3115/1073083.1073135.

[221] Kishore Papineni et al. *BLEU: a Method for Automatic Evaluation of Machine Translation*. Oct. 2002. DOI: 10.3115/1073083.1073135.

[222] Kishore Papineni et al. "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics. 2002, pp. 311–318.

[223] Thomas Fox Parry. *The Death of a Gig Worker*. June 1, 2018. URL: https://www.theatlantic.com/technology/archive/2018/06/gig-economy-death/561302.

[224] Devex Partnerships. *Q/A: Rwanda's radiology problem gets a startup solution*. 2021. URL: https://www.devex.com/news/sponsored/q-a-rwanda-s-radiology-problem-gets-a-startup-solution-101507.

[225] Samir Passi and Solon Barocas. "Problem Formulation and Fairness". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 39–48. ISBN: 9781450361255. DOI: 10.1145/3287560.3287567. URL: https://doi.org/10.1145/3287560.3287567.

[226] Adam Paszke et al. *Automatic differentiation in pytorch*. 2017.

[227] John Pavlopoulos et al. "Diagnostic Captioning: A Survey". In: *CoRR* abs/2101.07299 (2021). arXiv: 2101.07299. URL: https://arxiv.org/abs/2101.07299.

[228] Jay Peters. *Whole Foods is reportedly using a heat map to track stores at risk of unionization*. Apr. 20, 2020. URL: https://www.theverge.com/2020/4/20/21228324/amazon-whole-foods-unionization-heat-map-union.

[229] Sundar Pichai. *AI at Google: Our principles*. June 2018. URL: https://blog.google/technology/ai/ai-principles.

[230] Peng Qi et al. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, July 2020, pp. 101–108. DOI: 10.18653/v1/2020.acl-demos.14. URL: https://aclanthology.org/2020.acl-demos.14.

[231] Manish Raghavan et al. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 469–481. ISBN: 9781450369367. DOI: 10.1145/3351095.3372828. URL: https://doi.org/10.1145/3351095.3372828.

[232] Aniruddh Raghu et al. "Continuous State-Space Models for Optimal Sepsis Treatment: a Deep Reinforcement Learning Approach". In: *Proceedings of the 2nd Machine Learning for Healthcare Conference*. Ed. by Finale Doshi-Velez

et al. Vol. 68. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 147–163. URL: `https://proceedings.mlr.press/v68/raghu17a.html`.

[233] Sai Rajeswar et al. “Adversarial generation of natural language”. In: *arXiv preprint arXiv:1705.10929* (2017).

[234] Inioluwa Deborah Raji et al. “Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 33–44. ISBN: 9781450369367. DOI: `10.1145/3351095.3372873`. URL: `https://doi.org/10.1145/3351095.3372873`.

[235] Inioluwa Deborah Raji et al. “Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 33–44. ISBN: 9781450369367. DOI: `10.1145/3351095.3372873`. URL: `https://doi.org/10.1145/3351095.3372873`.

[236] Arun Rajkumar and Shivani Agarwal. “A Differentially Private Stochastic Gradient Descent Algorithm for Multiparty Classification”. In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Neil D. Lawrence and Mark Girolami. Vol. 22. Proceedings of Machine Learning Research. La Palma, Canary Islands: PMLR, Apr. 2012, pp. 933–941. URL: `https://proceedings.mlr.press/v22/rajkumar12.html`.

[237] Pranav Rajpurkar et al. “CheXnet: Radiologist-level pneumonia detection on chest X-rays with deep learning”. In: *arXiv preprint arXiv:1711.05225* (2017).

[238] Radim Rehurek and Petr Sojka. “Software framework for topic modelling with large corpora”. In: *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer. 2010.

[239] Ehud Reiter. “A Structured Review of the Validity of BLEU”. In: *Computational Linguistics* 44.3 (2018), pp. 393–401.

[240] Steven J Rennie et al. "Self-critical sequence training for image captioning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7008–7024.

[241] Steven J. Rennie et al. "Self-Critical Sequence Training for Image Captioning". In: *CVPR*. 2017.

[242] Manoel Horta Ribeiro et al. "Auditing Radicalization Pathways on YouTube". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 131–141. ISBN: 9781450369367. DOI: 10.1145/3351095.3372879. URL: https://doi.org/10.1145/3351095.3372879.

[243] Chris Mills Rodrigo. *Bipartisan senators call for investigation of popular fertility app*. 2020. URL: https://thehill.com/policy/technology/514169-bipartisan-senators-call-for-investigation-of-popular-fertility-app.

[244] Salvador Rodriguez. *Salesforce Says It Won't Build a Muslim Registry, Joining Apple, Facebook, Google, Others*. Dec. 19, 2016. URL: https://www.inc.com/salvador-rodriguez/salesforce-muslim-registry.html.

[245] Mica Rosenberg and Julia Edwards Ainsley. *Immigration hardliner says Trump team preparing plans for wall, mulling Muslim registry*. Nov. 16, 2016. URL: https://www.reuters.com/article/us-usa-trump-immigration-idUSKBN13B05C.

[246] Andrew B. Rosenkrantz, Danny R. Hughes, and Richard Duszak. "The U.S. Radiologist Workforce: An Analysis of Temporal and Geographic Variation by Using Large National Datasets". In: *Radiology* 279.1 (2016). PMID: 26509294, pp. 175–184. DOI: 10.1148/radiol.2015150921. eprint: https://doi.org/10.1148/radiol.2015150921. URL: https://doi.org/10.1148/radiol.2015150921.

[247] David A. Rosman et al. "Imaging in the land of 1000 hills: Rwanda radiology country report". In: *The Journal of Global Radiology* 1 (Mar. 2015). DOI: 10.7191/jgr.2015.1004. URL: https://rad-aid.org/wp-content/uploads/Imaging-in-the-Land-of-1000-Hills_-Rwanda-Radiology-Country-Repor.pdf.

[248] Deb Roy, Eugene Yi, and Russell Stevens. *Measuring The Health of Our Public Conversations*. Mar. 2018. URL: https://medium.com/cortico/measuring-the-health-of-our-public-conversations-d08d8d44f278.

[249] Fidel Rubagumya et al. "State of Cancer Control in Rwanda: Past, Present, and Future Opportunities". eng. In: *JCO global oncology* 6 (July 2020). PMC7392739[pmcid], pp. 1171–1177. ISSN: 2687-8941. DOI: 10.1200/GO.20.00281. URL: https://doi.org/10.1200/GO.20.00281.

[250] Shems Saleh et al. "Clinical Collabsheets: 53 Questions to Guide a Clinical Collaboration". In: *Proceedings of the 5th Machine Learning for Healthcare Conference*. Vol. 126. Proceedings of Machine Learning Research. PMLR, Aug. 2020.

[251] Tabinda Sarwar et al. "The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges". In: *ACM Comput. Surv.* 55.2 (Jan. 2022). ISSN: 0360-0300. DOI: 10.1145/3490234. URL: https://doi.org/10.1145/3490234.

[252] Noam Scheiber and Daisuke Wakabayashi. *Google Hires Firm Known for Anti-Union Efforts*. Nov. 20, 2019. URL: https://www.nytimes.com/2019/11/20/technology/Google-union-consultant.html.

[253] Dylan Scott. *Kamala Harris's plan to dramatically increase teacher salaries, explained*. 2019. URL: https://www.vox.com/policy-and-politics/2019/3/26/18280734/kamala-harris-2020-election-policies-teachers-salaries.

[254] Andrew D. Selbst et al. “Fairness and Abstraction in Sociotechnical Systems”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 59–68. ISBN: 9781450361255. DOI: 10.1145/3287560.3287598. URL: https://doi.org/10.1145/3287560.3287598.

[255] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. “BLEURT: Learning robust metrics for text generation”. In: *arXiv preprint arXiv:2004.04696* (2020).

[256] Mark Sendak et al. “"The Human Body is a Black Box": Supporting Clinical Decision-Making with Deep Learning”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 99–109. ISBN: 9781450369367. DOI: 10.1145/3351095.3372827. URL: https://doi.org/10.1145/3351095.3372827.

[257] Laleh Seyyed-Kalantari et al. “CheXclusion: Fairness gaps in deep chest X-ray classifiers”. In: *Pacific Symposium on Biocomputing*. 2021.

[258] Scott Shane and Daisuke Wakabayashi. *‘The Business of War’: Google Employees Protest Work for the Pentagon*. Apr. 4, 2018. URL: https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html.

[259] G Sharp and T.C. Schelling. *The Politics of Nonviolent Action*. Extending horizons books pt. 2. P. Sargent Publisher, 1973. ISBN: 9780875580685. URL: https://books.google.com/books?id=gA0XAAAAIAAJ.

[260] Desmond John Sheridan and Desmond G. Julian. “Achievements and Limitations of Evidence-Based Medicine.” In: *Journal of the American College of Cardiology* 68 2 (2016), pp. 204–13.

[261] Reza Shokri et al. “Membership Inference Attacks Against Machine Learning Models”. In: *2017 IEEE Symposium on Security and Privacy (SP)*. 2017, pp. 3–18. DOI: 10.1109/SP.2017.41.

[262] Cynthia Silva. *Top social media platforms 'unsafe' for LGBTQ users, report finds*. 2021. URL: https://www.nbcnews.com/nbc-out/out-news/top-social-media-platforms-unsafe-lgbtq-users-report-finds-rcna889.

[263] Cynthia Silva. *Top social media platforms 'unsafe' for LGBTQ users, report finds*. May 11, 2021. URL: https://www.nbcnews.com/nbc-out/out-news/top-social-media-platforms-unsafe-lgbtq-users-report-finds-rcna889.

[264] Ryan Singel. *Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims*. Dec. 17, 2009. URL: https://www.wired.com/2009/12/netflix-privacy-lawsuit.

[265] Natasha Singer and Cade Metz. *Many Facial-Recognition Systems Are Biased, Says U.S. Study*. 2019. URL: https://www.nytimes.com/2019/12/19/technology/facial-recognition-bias.html.

[266] Alexander K. Smith, Roger B. Davis, and Eric L. Krakauer. "Differences in the quality of the patient-physician relationship among terminally ill African-American and white patients: impact on advance care planning and treatment preferences". eng. In: *Journal of general internal medicine* 22.11 (Nov. 2007), pp. 1579–1582. ISSN: 1525-1497. DOI: 10.1007/s11606-007-0370-6. URL: https://pubmed.ncbi.nlm.nih.gov/17879120.

[267] Alexander K. Smith et al. "Racial and ethnic differences in advance care planning among patients with cancer: impact of terminal illness acknowledgment, religiousness, and treatment preferences". eng. In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 26.25 (Sept. 2008). 18757326[pmid], pp. 4131–4137. ISSN: 1527-7755. DOI: 10.1200/JCO.2007.14.8452. URL: https://pubmed.ncbi.nlm.nih.gov/18757326.

[268] Meg Sommerfeld. *Partnering for Compensation Reform: Collaborations between Union and District Leadership in Four School Systems*. 2011. URL: https://cdn.americanprogress.org/wp-content/uploads/issues/2011/

06/pdf/union_district.pdf?_ga=2.135672244.46324891.1631253156-235729095.1631253156.

[269] Stuart N. Soroka and Christopher Wlezien. “The Thermostatic Model”. In: *Degrees of Democracy: Politics, Public Opinion, and Policy*. Cambridge University Press, 2009, pp. 22–42. DOI: 10.1017/CBO9780511804908.004.

[270] Preethi Srinivasan et al. “Hierarchical X-Ray Report Generation via Pathology Tags and Multi Head Attention”. In: *Computer Vision – ACCV 2020*. Ed. by Hiroshi Ishikawa et al. Cham: Springer International Publishing, 2021, pp. 600–616. ISBN: 978-3-030-69541-5.

[271] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. “Synthetic Data – Anonymisation Groundhog Day”. In: *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/stadler.

[272] "Mayo Clinic Staff". *Living wills and advance directives for medical decisions*. 2020. URL: https://www.mayoclinic.org/healthy-lifestyle/consumer-health/in-depth/living-wills/art-20046303.

[273] Maria J. Stephan and Erica Chenoweth. “Why Civil Resistance Works: The Strategic Logic of Nonviolent Conflict”. In: *International Security* 33.1 (2008), pp. 7–44. ISSN: 01622889, 15314804. URL: http://www.jstor.org/stable/40207100.

[274] Charlotte Stix. *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*. Tech. rep. European Commission High-Level Expert Group on AI (AI HLEG)), 2020. URL: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[275] Leah Cardamore Stokes. *Short Circuiting Policy: Interest Groups and the Battle Over Clean Energy and Climate Policy in the American States*. Oxford University Press, 2020.

[276] Cathie Sudlow et al. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age". In: *PLOS Medicine* 12.3 (Mar. 2015), pp. 1–10. DOI: 10.1371/journal.pmed.1001779. URL: https://doi.org/10.1371/journal.pmed.1001779.

[277] Elior Sulem, Omri Abend, and Ari Rappoport. "Bleu is not suitable for the evaluation of text simplification". In: *arXiv preprint arXiv:1810.05995* (2018).

[278] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems*. 2014, pp. 3104–3112.

[279] Latanya Sweeney. "<i>K</i>-Anonymity: A Model for Protecting Privacy". In: *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10.5 (Oct. 2002), pp. 557–570. ISSN: 0218-4885. DOI: 10.1142/S0218488502001648. URL: https://doi.org/10.1142/S0218488502001648.

[280] Tanveer Syeda-Mahmood et al. "Chest X-Ray Report Generation Through Fine-Grained Label Learning". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Ed. by Anne L. Martel et al. Cham: Springer International Publishing, 2020, pp. 561–571. ISBN: 978-3-030-59713-9.

[281] Jun Tang et al. *Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12*. 2017. DOI: 10.48550/ARXIV.1709.02753. URL: https://arxiv.org/abs/1709.02753.

[282] Collective Action in Tech archive. 2022. URL: https://data.collectiveaction.tech.

[283] Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness Revised and Expanded Edition*. Penguin Books, 2009. ISBN: 9780143115267.

[284] Big tech companies back away from selling facial recognition to police. That's progress. *Rebecca Heilweil*. June 11, 2020. URL: https://www.vox.com/

recode/2020/6/10/21287194/amazon-microsoft-ibm-facial-recognition-moratorium-police.

[285] "The "All of Us" Research Program". In: *New England Journal of Medicine* 381.7 (2019), pp. 668–676. DOI: 10.1056/NEJMsr1809937. URL: https://doi.org/10.1056/NEJMsr1809937.

[286] *Too Much Medicine*. https://www.bmj.com/too-much-medicine. Accessed: 2022-05-12.

[287] Kevin Truong. *The Open Source Community Is Calling on Github to 'Drop ICE'*. July 20, 2020. URL: https://www.vice.com/en/article/m7jpgy/open-source-community-changing-github-avatars-drop-ice.

[288] Agnes Uhereczky. *Old Meets New: Google Employees Just Formed A Union.* Jan. 6, 2021. URL: https://www.forbes.com/sites/agnesuhereczky/2021/01/06/old-meets-new-google-employees-just-formed-a-union.

[289] R. Vedantam, C. L. Zitnick, and D. Parikh. "CIDEr: Consensus-based image description evaluation". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015, pp. 4566–4575. DOI: 10.1109/CVPR.2015.7299087.

[290] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4566–4575.

[291] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. "CIDEr: Consensus-based image description evaluation." In: *CVPR*. IEEE Computer Society, 2015, pp. 4566–4575. ISBN: 978-1-4673-6964-0. URL: http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html#VedantamZP15.

[292] Nicholas Vincent et al. "Data Leverage: A Framework for Empowering the Public in Its Relationship with Technology Companies". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 215–

227. ISBN: 9781450383097. DOI: 10.1145/3442188.3445885. URL: https://doi.org/10.1145/3442188.3445885.

[293] Oriol Vinyals et al. "Show and tell: A neural image caption generator". In: *Computer Vision and Pattern Recognition*. 2015. URL: http://arxiv.org/abs/1411.4555.

[294] Robert Wachter. *The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age*. McGraw-Hill Education, 2015. ISBN: 978-1260019605.

[295] Kurt Wagner. *This is how Facebook collects data on you even if you don't have an account*. 2018. URL: https://www.vox.com/2018/4/20/17254312/facebook-shadow-profiles-data-collection-non-users-mark-zuckerberg.

[296] Kurt Wagner. *This is how Facebook collects data on you even if you don't have an account*. Apr. 20, 2018. URL: https://www.vox.com/2018/4/20/17254312/facebook-shadow-profiles-data-collection-non-users-mark-zuckerberg.

[297] Daisuke Wakabayashi et al. *Google Walkout: Employees Stage Protest Over Handling of Sexual Harassment*. Nov. 1, 2018. URL: https://www.nytimes.com/2018/11/01/technology/google-walkout-sexual-harassment.html.

[298] Xiang Wang, David Sontag, and Fei Wang. "Unsupervised Learning of Disease Progression Models". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '14. New York, New York, USA: Association for Computing Machinery, 2014, pp. 85–94. ISBN: 9781450329569. DOI: 10.1145/2623330.2623754. URL: https://doi.org/10.1145/2623330.2623754.

[299] Xiaosong Wang et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2017, pp. 3462–3471.

[300] Xiaosong Wang et al. "TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 9049–9058.

[301] Harriet Washington. *Medical Apartheid: The Dark History of Medical Experimentation on Black Americans from Colonial Times to the Present*. Doubleday, 2007. ISBN: 978-0385509930.

[302] Ryan Webster et al. "This Person (Probably) Exists. Identity Membership Attacks Against GAN Generated Faces". In: *ArXiv* abs/2107.06018 (2021).

[303] Maranke Wieringa. "What to Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 1–18. ISBN: 9781450369367. DOI: `10.1145/3351095.3372833`. URL: `https://doi.org/10.1145/3351095.3372833`.

[304] David R. Williams. "Race, Socioeconomic Status, and Health The Added Effects of Racism and Discrimination". In: *Annals of the New York Academy of Sciences* 896.1 (1999), pp. 173–188. ISSN: 1749-6632. DOI: `10.1111/j.1749-6632.1999.tb08114.x`. URL: `http://dx.doi.org/10.1111/j.1749-6632.1999.tb08114.x`.

[305] Ronald J Williams. "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine learning* 8.3-4 (1992), pp. 229–256.

[306] Michael Wines. *Deceased G.O.P. Strategist's Hard Drives Reveal New Details on the Census Citizenship Question*. 2019. URL: `https://www.nytimes.com/2019/05/30/us/census-citizenship-question-hofeller.html`.

[307] Langdon Winner. "Do Artifacts Have Politics?" In: *The MIT Press*. 1980. URL: `http://www.jstor.org/stable/20024652`.

[308] Sam Wiseman, Stuart M Shieber, and Alexander M Rush. "Challenges in data-to-document generation". In: *arXiv preprint arXiv:1707.08052* (2017).

[309] Ziang Xie. "Neural text generation: A practical guide". In: *arXiv preprint arXiv:1711.09534* (2017).

[310] Jia Xu et al. "Differentially Private Histogram Publication". In: *2012 IEEE 28th International Conference on Data Engineering*. 2012, pp. 32–43. DOI: 10.1109/ICDE.2012.48.

[311] Kelvin Xu et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". en. In: *arXiv:1502.03044 [cs]* (Feb. 2015). arXiv: 1502.03044. URL: http://arxiv.org/abs/1502.03044 (visited on 03/13/2019).

[312] Kelvin Xu et al. "Show, attend and tell: Neural image caption generation with visual attention". In: *International conference on machine learning*. 2015, pp. 2048–2057.

[313] Wenting Xu et al. "Reinforced Medical Report Generation with X-Linear Attention and Repetition Penalty". In: *CoRR* abs/2011.07680 (2020). arXiv: 2011.07680. URL: https://arxiv.org/abs/2011.07680.

[314] Kuldeep N. Yadav et al. "Approximately One In Three US Adults Completes Any Type Of Advance Directive For End-Of-Life Care". In: *Health Affairs* 36.7 (2017), pp. 1244–1251. DOI: 10.1377/hlthaff.2017.0175. eprint: https://doi.org/10.1377/hlthaff.2017.0175. URL: https://doi.org/10.1377/hlthaff.2017.0175.

[315] Adam Yala et al. "Toward robust mammography-based models for breast cancer risk". In: *Science Translational Medicine* 13.578 (2021), eaba4373.

[316] Eddie Yang and Margaret E. Roberts. "Censorship of Online Encyclopedias: Implications for NLP Models". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 537–548. ISBN: 9781450383097.

DOI: 10.1145/3442188.3445916. URL: https://doi.org/10.1145/3442188.3445916.

[317] Ting Yao et al. “Boosting image captioning with attributes”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 4894–4902.

[318] Matthew Yglesias. *They deliberately put errors in the Census*. Aug. 16, 2021. URL: https://www.slowboring.com/p/census-privacy.

[319] Tianyi Zhang et al. “Bertscore: Evaluating text generation with bert”. In: *arXiv preprint arXiv:1904.09675* (2019).

[320] Yixiao Zhang et al. “When Radiology Report Generation Meets Knowledge Graph”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.07 (Apr. 2020), pp. 12910–12917. DOI: 10.1609/aaai.v34i07.6989. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6989.

[321] Yuhao Zhang et al. “Learning to summarize radiology findings”. In: *arXiv preprint arXiv:1809.04698* (2018).