Negotiating The Future: Leveraging Socio-technical Narratives to Engage Multiple Voices in the Ethics of our Future

by

Michael Warren Skirpan

B.S., University of Pittsburgh, 2010

B.Phil, University of Pittsburgh, 2010

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2017

This thesis entitled:
Negotiating The Future: Leveraging Socio-technical Narratives to Engage Multiple Voices in the Ethics of our Future
written by Michael Warren Skirpan
has been approved for the Department of Computer Science

Prof. Tom Yeh

Prof. Clayton Lewis

Prof. Casey Fiesler

Date

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

IRB protocol #16-0130 and 17-0260

Michael Warren Skirpan, (Ph.D., Computer Science)

Negotiating The Future: Leveraging Socio-technical Narratives to Engage Multiple Voices in the Ethics of our Future

Thesis directed by Prof. Tom Yeh

As technology mediates more of our lives and leverages more of our personal data, ethical, legal, and policy questions gain more relevance and place pressure on our institutions and societal norms. This thesis looks at how the use of robust socio-technical narratives can act as a space for considering the correspondence between technical choices and social consequences of technology. Given the diversity of stakeholders who shape and are impacted by technology – such as user communities, lawyers, engineers, students, policymakers – it is critical we develop communication strategies and mediums for negotiating the sticky ethical and social questions relevant to technology development. This thesis explores several research areas where narrative acts as a tool for discussion, clarification, and negotiation about future uses of emerging technologies. The research work of this dissertation leverages narrative theoretically to support a policy framework for future regulation of data-driven technology. It further looks at how narrative can facilitate ethics discussion in the CS classroom, clarify and raise ethical issues for debate by engineers, and create a space for experts and non-experts to hold discourse and express values around technology ethics and policy. This work draws from several literatures including HCI topics such design fiction and user enactments, risk perception, policy, philosophy, law, and art.

Dedication

To my love Jackie, all the friends who propelled my thinking over the years, and all those who doubted that a guy with a hardline ethical agenda could get anywhere.

Acknowledgements

Every idea and result in this thesis became a reality due to a culmination of support. I start by acknowledging my wife and life partner, Jackie (Cameron) Skirpan, who has helped me turn my ideas into reality as my co-producer, co-author, and closest intellectual ally. My dear friend Patrick Cooper has helped feed me interesting ideas and been my smartest interlocutor for nearly a decade, there's no way I'd be here without him. Micha Gorelick, another close friend and the best programmer I have ever met, who has mentored me through the development of my CS career and helped me understand the technical details of everything in this thesis. Tom Yeh, my advisor, has been a pillar of support for my research and never questioned my abilities and took every idea I threw at him as exciting and serious. Clayton Lewis, a wonderful mentor and guide throughout my time at CU.

I also want to thank the rest of the committee - Paul Ohm, Casey Fiesler, and Tamara Sumner for their amazing support and words of wisdom. My roommates in Colorado - Simone Hyater-Adams and Julie Cafarella - who constantly gave me their ears and minds. Willie Costello my co-author and old philosophy friend who helped me ground my theoretical work. My home lab at Colorado: Sikuli Lab. Then the cast of amazing people who can't go unnamed due to their contributions to my life and intellect while doing this work: Nathan Cochran, Jon Weisberg, Chris Stokum, Isaac Gaylord, Brittany Kos, Sam Molnar, Travis Alvarez, Luke Underwood, and Nathan Zimmerman. And some of my great teachers: John Black, Liz Bradley, Ronald Judy, and John McDowell. Plus the many others at conferences, meet-ups, bars, and coffee shops who played a role in shaping my ideas.

Contents

Chapter

1 Introduction **1**

1.1 Technology as Tragedy or Hope 1

1.1.1 Notifications for Whom?: A Framing Example 4

1.2 The Immanent Need for Data and Machine Ethics 8

1.3 The Role of Narrative in Ethics Research 13

1.3.1 Background of Narrative in HCI and Ethics Research 14

1.3.2 Speculative Ethics Using Templates, Targets, and Boundaries 21

1.3.3 Interrogating the Narrative: Ethics Research and Dialogue In Practice 28

1.4 Outline of the Work Presented 33

2 The Challenge Ahead: Why We Need New Ethical Thinking in Technology that Leverages Multiple Stakeholders **36**

2.1 A Survey of Data Ethics: Problems New and Old 36

2.1.1 Prologue 36

2.1.2 Introduction 36

2.1.3 Privacy: A Right and a Preference 41

2.1.4 Discrimination: Reorienting the Problem to the Machine 51

2.1.5 Online Research, Consent, and User Attitudes 57

2.1.6 Algorithmic Impact 63

2.1.7 The Future of (Data) Ethics 70
2.2 The Authority of "Fair" in Machine Learning 76
2.2.1 Prologue 76
2.2.2 Introduction 76
2.2.3 The Construction of "Fairness" 77
2.2.4 Current State of Fair ML 81
2.2.5 Concluding Remarks 85

3 Framing Policy **86**
3.1 Coding For Respect 86
3.1.1 Prologue 86
3.1.2 Introduction 87
3.1.3 The concept of respect 89
3.1.4 An analysis of machine action 91
3.1.5 Respect in machine action 96
3.1.6 Concluding Remarks 107

4 Ethical Thinking within Computer Science **109**
4.1 Ad Empathy: A Design Fiction 109
4.1.1 Prologue 109
4.1.2 Product Introduction 109
4.1.3 Getting Started 110
4.1.4 API Resources 110
4.1.5 How Does It Work? 112
4.1.6 Example Use 114
4.1.7 Appropriate Use of Ad Empathy 115
4.1.8 Author's Statement 116

4.2 Ethics Education in Context: A Case Study of Novel Ethics Activities for the CS Classroom . . . 118
4.2.1 Prologue . . . 118
4.2.2 Introduction . . . 118
4.2.3 Literature Review . . . 120
4.2.4 Course Overview . . . 121
4.2.5 Results . . . 125
4.2.6 Discussion . . . 129
4.2.7 Concluding Remarks . . . 131
4.3 Quantified Self: An Interdisciplinary Immersive Theater Project Supporting a Collaborative Learning Environment for CS Ethics . . . 131
4.3.1 Prologue . . . 131
4.3.2 Introduction . . . 132
4.3.3 Prior Work and Motivation . . . 133
4.3.4 Approach . . . 135
4.3.5 Evaluation . . . 140
4.3.6 Discussion . . . 143
4.3.7 Conclusion . . . 144
4.4 Designing a Moral Compass for Computer Vision Using Speculative Analysis . . . 145
4.4.1 Prologue . . . 145
4.4.2 Introduction . . . 145
4.4.3 Categorizing Risk Factors in Computer Vision . . . 148
4.4.4 Scenario Evaluation . . . 151
4.4.5 Narrative Case Studies . . . 158
4.4.6 Conclusion & Future Work . . . 162

5 Technology Ethics in the Public Sphere **166**

5.1 What's at Stake: Characterizing Risk Perceptions of Emerging Technologies 166
5.1.1 Prologue 166
5.1.2 Introduction 166
5.1.3 Related Work 168
5.1.4 methods 170
5.2 Results 173
5.2.1 Properties of Population Sample 174
5.2.2 Risk Ranking and Scoring 175
5.2.3 Psychological Factors of Risk 177
5.2.4 Worst Case Scenarios 180
5.2.5 Additional Risks 183
5.2.6 Discussion 184
5.2.7 Risk-Sensitive Design 187
5.2.8 Conclusion 192
5.3 More Than a Show: Using Personalized Immersive Theater to Educate and Engage the Public in Technology Ethics 193
5.3.1 Prologue 193
5.3.2 Introduction 193
5.3.3 Related Work 194
5.3.4 Design Space 196
5.3.5 Approach and Methods 199
5.3.6 Results 209
5.3.7 Discussion 217
5.3.8 Conclusion 220

6 Conclusion **221**

6.1 Reflections on the Work 221

6.1.1 The Growing Complexity of the Problem 221

6.1.2 The Promises of Narrative and What's Comes Next 223

6.2 Reader's Guide 226

Bibliography **227**

Tables

Table

3.1 The components of machine action 92

3.2 Cases in our matrix of machine action 97

3.3 Norms of respect 106

3.4 Questions of respect 106

4.1 Twenty-Two Risky Scenarios Used for Analysis 165

5.1 Scenario Types, Risks, and Example Text from Respondents' Open-Ended Worst-Case Scenario Response 182

5.2 Selected exhibits to expose audience to certain personal data use issues. N denotes the number of users an exhibit is designed for. 206

Figures

Figure

1.1 A graph originally drawn by Joseph Voros, redrawn here by Stuart Candy, that delimits the space of the future into the possible, plausible, probable, and preferable. 20

1.2 The network of templates, potentially shaped by targets or boundaries, that go into forming a socio-technical narrative. 24

1.3 This graphic represents the evaluative questions that can be applied to a socio-technical narrative. Take note that the different layers may be best approached by different stakeholders; though the goal is that the lowest layer underpins them all and can be discovered within any expert or non-expert inquiry. 29

1.4 Here we have the broad representation of designing, creating, and evaluating a socio-technical narrative within my framing. 32

4.1 Likelihood categories of possible futures. 155

4.2 Uncertainty vs. Severity for 22 CV Risk Scenarios 157

4.3 Deeply Learned Bias (Scenario 10) 159

4.4 The Cameras Attack (Scenario 21) 160

5.1 Average comparative risk ranking by non-experts vs experts where items with significant differences (p¡.05 for two-tailed t-test) are highlighted. 176

5.2 Risk perception by experts vs non-experts for 18 technologies. 178

5.3 Comparing the psychological factors regarding the filter bubble (orange) vs research without consent (blue) as perceived by non-experts (dotted) vs by experts (solid). . 181

5.4 Risk-Sensitive Design proposes that risk mitigation strategies should be informed by the difference between public and experts' risk perception and the degree to which such difference is acceptable ($E - M$). 189

5.5 Our graphical model of the affordances offered by a tech/art project's design structure.198

5.6 The goal of Quantified Self is to afford a balanced experience across all dimensions of our heuristic (1). We came close to meeting this goal (2). Below are examples of other projects using our heuristic for considering the affordances of different designs (a-f). 200

5.7 Time line for Quantified Self production. 202

Chapter 1

Introduction

"I have a foreboding of an America in my children's or grandchildren's time–when the United States is a service or information economy; when nearly all the key manufacturing industries have slipped away to other countries; when awesome technological powers are in the hands of a very few and no one representing the public interest can even grasp the issues; when the people have lost the ability to see their own agendas or knowledgeably question those in authority; when, clutching our crystals and nervously consulting our horoscopes, our critical faculties in decline, unable to distinguish between what feels good and what's true, we slide, almost without noticing, back into superstition and darkness."

-Carl Sagan, The Demon-Haunted World

1.1 Technology as Tragedy or Hope

As humankind treads further into the Twenty-First Century, it is becoming overwhelmingly apparent that the fate of our species and the evolution of technology are deeply bound together. The challenges we face as a society and a planet are compounding day after day. A short list may include: rise in our global temperature, destruction of scarce natural resources, increasing gaps between rich and poor, looming potentials of nuclear war, lacking and unequal access to medicine and education, continuance of racism and sexism, and, here in the United States, a severe inability to communicate across cultural, generational, and ideological lines. Nonetheless, one piece of this complex puzzle that retains hope and possibility is the promise of technology. The potential of the

yet-to-be-found invention or idea. Can we innovate our way out of the many crises humanity faces?

Technology, thought of as the evolution of our human capacity to solve problems, is not merely a set of external tools we leverage. Rather, technology sets the horizon of human capabilities. It is fundamental to how we conceptualize ourselves. The idea of technology itself is adopted into our collective psychology; at least for those privileged to have access. Those of us with technological access, strategize our lives with the presumption that technologies are at our disposal. Thusly, our capacity to mitigate the mounting problems of our world come down to what powers we are able to wield for solving them. Not just now, but in the future. However, the implication that technology can be wielded for good is bolstered by the specious assumption that we are in control of the powers technology bestows. Assuming we are in control of technology may appear true at face value. Though if you take a deeper look at a complex technology like the search engine, one may ask, "Is the algorithm a tool for me to give inputs and get information or does the algorithm instrumentalize me by ordering information to elicit my click?" Meditating on a severe case like that of Dylann Roof [11] should at least put a pause on any immediate conclusion that the search engine has solved the problem of *quickly getting factual information*. What instead we may wonder is, "Under what framing of the problem has technology provided the solution?" and then in reflection, "Is that the best framing of the problem?"

Increasingly we are seeing that the emergence of new technologies alone will not save us from our hardest societal problems. In fact, we are now facing vast evidence that, when deployed naively, technology can cause more harm than good or alternatively, offer benefits to only the select few who can afford it. Automation and AI are continually overtaking the job market, accruing wealth into fewer hands as our physical, and now mental, labor recedes slowly toward obsolescence [71]. Just as quickly as social media availed to us the possibility of widespread communication and human relationships overcoming physical barriers of distance, it admonished us by allowing the rapid spread of misinformation and the fracturing of the public sphere through filter bubbles [241, 123]. And now too, as we hoped machines would optimize away our human biases, we are seeing that machine intelligence is capable of adopting the very prejudices we are fighting against [42, 39, 27].

These exemplary problems are only leaves of a larger tree of problems related to technology. From the dwindling meaning of the right to privacy to the reality of autonomous weapons to the harsh fact that the precious metals that make up our technology are sourced through extractive and exploitative means, it is unclear what the cost-benefit balance sheet for technology truly is. Which is not to say that we must slam the brakes, but instead recognize that if technology is to solve anything for us, it must be designed, implemented, and deployed with forethought and purpose.

It is here, stating the desire for technology to be developed with forethought and purpose, that we enter into the motivations for the work that follows. This thesis is an exploration of what I will argue is a central issue for our future: How do we develop technology in the most ethical way possible? Unpacked, this question opens up some of the most difficult and important sub-problems technologists, policymakers, and the public face. Problems such as, What testing should be done on a feature prior to release? How do we evaluate the social impacts a technology will have on its users? How do we design technologies that align with the values of its users? Who is to blame when a technology allows for severe, unintended harm? How do we leverage Big Data while respecting privacy? What is informed consent and how do we achieve it at scale? How do we train AI systems that do not inherit unwanted bias? How do we design machine intelligent systems that are auditable and transparent? What rights should users have with respect to entities that license and use their data? How do we inform the public about technology so that they can make decisions in their best interests? What does it mean to instrumentalize humans to the purposes of machines? How does a user meaningfully disagree with a decision made by a machine?

These questions are not mere philosophical meanderings. They each attach very directly to choices technologists are making *today*. Yet, these questions have no straightforward answers. Not only do they require the treacherous value-laden judgments demanded by ethical inquiry, but they further demand our having insight into the unknown, namely, the future. Put another way, thinking about the ethics of technology carries with it a two-fold challenge. (1) we have to translate the space of subjective values into the space of technological choices and (2) our ethical analysis must look beyond what we assume based on our current situation to account for the impacts that

may occur after the technology is released.

Skipping to the punchline, the central theme of this dissertations is as follows: crafting socio-technical narratives to represent the future of technology is our most powerful human tool for meeting the challenges of (1) and (2) above. Narratives, sometimes referred to as scenarios or vignettes, have a unique power to blend the world of technical artifacts with that of human values and relationships. What I will argue in this introduction and provide evidence for through the chapters, is that thoughtful narrative can represent socio-technical problems such that multiple stakeholders, of varying backgrounds, can meaningful consider the ethical implications of a technology. Further, I will argue that narrative can be used to clarify our assumptions, identify unforeseen problems, and present alternatives about the promise and danger of new technologies. Before getting into the background of this work and discussing the limit of these claims, let's pause to consider a case study. To clarify how exactly it is that an ambiguous and broad tool such as **narrative** could be useful for responding to the ethical challenges listed under (1) and (2), allow me to consider the technical and design problem of the notification.

1.1.1 Notifications for Whom?: A Framing Example

Smart phones and computers have the responsibility of managing vast amounts of information for our daily lives. Two important problems they attempt to solve is telling a user when new information is available or reminding them of information when it is pertinent. Relative to these problems, we have notifications. Anyone with a smart phone likely knows, notifications are almost unavoidable features given the number of applications running in parallel and the need to allow new information from those applications to be received asynchronously so the user can do one task without entirely forgetting another.

Immediately a designer can enumerate questions such as: How disruptive should a notification be? Where should we put them on a screen? For how long? Which notifications should default to being off or hidden? What privacy constraints should be available for the user? The list goes on. In tandem, engineers designing operating systems have to ask: What permissions should an appli-

cation receive to show notifications? When multiple notifications are available, how should they be ordered? Should some notifications be unable to be ignored? Again, the list of considerations is too long to exhaust. It should be noted, that this example abstracts away from other practical realities that would be considered such as, "How do we make sure users pay attention to what we want them to?" and "How can we make notifications most advantageous for our company and partner companies?" But for now, let's treat the situation as technical design problem unencumbered by organizational demands.

From a user vantage, we similarly have a set of considerations: How do I silence notifications? How do I prevent unwanted notifications? How do I make sure I do not violate privacy of my friends or colleagues with an overt notification? How do I make sure I do not miss something important to me? How can I make sure I do not forget something I must remember? How do I make sure the phone gets my attention? And, of course, each user has their own needs and priorities (some may not care about notifications at all).

Now, given this situation where a designer and an engineer are working together to solve notification problems for a user, I can ask some ethical questions of their choices. It strikes me as likely that some designers and engineers would balk at the idea that, beyond severe privacy concerns, there are many ethical questions to ask in this realm. Yet, I'll start with a fundamental question, "Does the system treat the user as an autonomous agent with its own goals and values?" This high-level question can then be unpacked into a number of sub questions. For example:

(1) What does your ordering algorithm optimize for?

(2) How does a user inform the system that they require a pause in notifications or change in the notification schema?

(3) Does your notification system treat all users the same?

(4) Is your system capable of subverting the users goals?

Let's consider each one of these briefly. (1) is asking about value alignment. It is likely (1)

is answered by an optimization for most commonly used applications, most clicked notifications, or a priority of applications by type (e.g., $alarms > email > socialmedia > newsarticles$). This is an attempt to codify the users values, or likely values, as observed through behavior. We would find it disturbing if the answer was, "We analyze what content is likely to make the user upset and show that first." Given that no user anticipates or feels they agreed to something like this, it would be egregious and unethical to utilize such an algorithm. (2) would likely be responded to by explaining that notifications can be paused or silenced through a menu interface or permission setting. (3) starts getting into more difficult territory conceptually. For the non-technical user, it may not really be obvious what this even means, but for the engineer, they may immediately think about personalization. Telling the user, "your notifications are personalized given what we assess of your behaviors and personality," may, to some extent, be the spurious response to dealing with (3) and likely alleviate a designer's concern due to a moral rationalization of consent. However, if we were to explain that, "it means men and women will likely get significantly different notifications," they may not be so excited. Now this question becomes one about fairness and discrimination; something the engineer implementing a personalized algorithm was possibly not thinking about.

Moving to perhaps the most challenging of the questions, (4), we might imagine that an engineer simply say, "of course not," or, "what does that even mean?" We would also imagine a user saying they obviously do not want this, but are not fully sure what it means. And it is here that we really start getting in muddy water. Leading up to (4), the previous questions carried some amount of intuition and we imagine the engineers would have reasonable explanations regarding making their best attempts to accommodate user values to the extent possible. While obviously concerning, (4) denies intuition and instead leaves one thinking, "well of course no one would try to do such a thing." However, a very clever engineer (or creative layperson) may read that question and think about some complex dynamics that could emerge where, in attempting to algorithmically optimize for a click, a user's cognitive abilities may, against their interest or desire, lose to a very intelligent notification system. Still it is almost difficult to imagine. Thus, let's try a narrative.

"Shawna has owned her phone for about a year. She mostly uses it for email, social media,

and reading the news, but also does some online shopping. Usually shopping on her phone is restricted to her commute to work. Her phone has permissions to show her notifications when sales are occurring on her favorite shopping applications. Her phone is also designed so that it shows notifications when they are most relevant. Until recently, this has been great for Shawna. When reading the news, her phone learns not to interrupt her with social media and when typing emails, it has learned not to bother her with news. Recently, Shawna looked through her bank account and realized that she has continued to spend more money on clothes over the past two months. After this realization, she began picking up on the fact that her shopping notifications were alerting her at new and odd times, often late in the evening, right after she was on social media. Now that she put the pieces together, she realized this pattern started after posting that she felt jealous of a friend of her's who is always showing off brand new outfits online. Unbeknownst to her, the smart notification system had picked up on the fact that Shawna was cognitively most disposed to click and buy after leaving certain behavioral data traces that were signals of jealousy and insecurity. Not fully understanding what happened, Shawna now is now feeling violated, a bit confused, and uncertain of what her phone is capable of doing."

Reading this, some people may immediately roll their eyes and say this is ridiculous and very unlikely. Others may raise their eyebrows and think about what pattern of inputs fed to an algorithm may actually create this scenario. In either case, we can very easily say that the algorithm, as used in this scenario, led to a cognitive manipulation effect that was negative for the user (and made the user's interests subordinate to its optimization). This case is motivated by a real marketing strategy (e.g., [231, 287]).

If we were to extend this case to other domains outside of online shopping, we can surely imagine that a similar manipulation could have even more chilling consequences. What's crucial is that the form of this story allows nearly anyone to understand the problem it raises. Thus, now we actually have a real space where (4) can be debated from a common example. E.g., Can an algorithm do that? Should an algorithm be able to do that? If an algorithm does that, does the person have a right to recourse? Is this really covered by consent? I will refrain from attempting

to answer how this narrative, question (4), or these sub-questions should be resolved. Though, I will simply state that engineers should work to make systems where this narrative can be avoided, or at least where a user could control the inputs being given to the algorithm so they can limit or fix such issues.

Ending this aside here, I hope this motivates the possibilities for why narrative can elucidate complicated ethical situations created by technology. Specifically, it supports the difficult translation of social values into technical systems and offers grounded visions around uncertain and hard to grasp future consequences.

1.2 The Immanent Need for Data and Machine Ethics

Having made grand statements about the dire need for ethics in technology and why narrative may support this process, let's go layer deeper into *why* these question are so pertinent at the present moment and *what* focus I am taking in this thesis among myriad relevant ethical topics. Scattered throughout the following chapters, dozens of examples will be discussed where the use of Big Data and Machine Intelligence cause unexpected and concerning impacts. Whether it is the rapid loss of privacy due to the heightened difficulty to anonymize [238, 250], increased ability to predict private information [96, 194] or the complex ways in which algorithms are able to shape and harm humans [147, 321]. All of these questions are ripe for investigation in the era of the "Fourth Paradigm" [165]. That is, as I will overview in the second chapter of this thesis, the rise of Big Data is changing the wiring of thought and humanity, and in turn pressing on our moral and legal ideas. The idea of the "Fourth Paradigm" is that the increased availability of data from online sources, sensors, synthetic simulations, etc means that our ability to track, model, and predict phenomena using no hypothesis, theory, or scientific method is the methodology of the day. And the critical point to recognize is that hype around an AI revolution [71] is intrinsically wed to the Big Data revolution.

In her talk, "Machine Intelligence Makes Human Morals More Important" [322], Zeynep Tufekci urges us to see that through the use of trained AI systems, we are beginning to offload our moral reasoning onto machines. Contentious and sensitive matters such as policing, prison

sentencing, and mental health diagnosis are now being encroached upon by machine-intelligent systems. Her warning to us is, "if we do not start debating and choosing moral guidelines, machines will dominate these spaces without our say." This rings harmonious with a prior warning argued by the famed computer scientist and cultural critic, Joseph Weizenbaum. In his seminal book, "Computer Power and Human Reason," Weizenbaum argues that humans make choices whereas computers make decisions [342]. He worries that there is a confusion around the capabilities of computer power that will neuter our interest in the normative, ethical, political nuance of human reason. In particular, he sees a future where we are unable to tease out the value-laden propositions of science and aim toward a society that upholds the our human values. His postulation is that this could happen due to the relinquishing of human judgment in favor of machines that can blindly make decisions for us.

Historians too have ruminated on this very concern. Prized historian Lewis Mumford, in his book "Technics and Civilization," specifically argues that humans have never fully found a purpose for machines in human society. Rather, our wanderlust to let machines recapture our old ideas in powerful and efficient ways distracted us from some of the fundamental social questions raised by machines in the early industrial era. He writes,

"The machine itself makes no demands and holds out no promises: it is the human spirit that makes demands and keeps promises. In order to reconquer the machine and subdue it to human purposes, one must first understand it and assimilate it. So far, we have embraced the machine without fully understanding it, or, like the weaker romantics, we have rejected the machine without first seeing how much of it we could intelligently assimilate." [234, p.6]

Again, the problem is seeded in similar soil. Blindly applying machine power to human problems without considering what we want to accomplish socially may allow humans to be nothing more than instrumental to a society of machines carrying out industrial demands. As if the horrific consequences of rapid top soil removal triggering the Dust Bowl [75] weren't evidence enough to tread with some care when harnessing fantastic new powers. The logic of "innovate first, think about it later," may swallow us whole before we can tame technology to our purpose. And a logic

of this kind is being applied all around us in relation markets around our Big Data, now housed and licensed by the largest online providers. Nearly any machine intelligence expert would agree that new paradigms, such as deep learning, give us the capacity to turn vast amounts of unstructured data into "accurate" models of arbitrary kind. So where exactly does human value fit into the equation of training machines on vast quantities of data?

Take DARPA's stance on AI [202], which sets out three primary categories: handcrafted knowledge, statistical learning, and contextual adaptation. Handcrafted knowledge is the earliest paradigm of AI where experts programmed pre-structured sets of rules to represent knowledge and then allow machines to process data using these structures. Statistical learning, the current paradigm, involves engineers leveraging statistical models that are trained on data. The fundamental idea is that once enough data passes through the system, the machine will learn how to separate features into desired classes required for decision-making and inference (ie, train a model on enough criminal profiles and it will "get good at" identifying a likely criminal). Think of it as a very slow version of how you train a child to do a task through trial and error. The upcoming paradigm, DARPA claims, is contextual adaptation, which uses explainable models to support contextual inquiry into how abstractions and inferences are being made. The goal of contextual adaptation is to allow machines to become even more human and identify *which model* is best suited to handle a particular input rather than the current single-model approach. Now we might ask, whether in the statistical learning or the contextual adaptation paradigm, Who decides how to restrain what data models can utilize and toward what goals? Who decides how a machine decides which context is most appropriate for a choice? Is the goal to resign the human and "let the data decide?" In alignment with Mumford's worries, Can we still promise humans the rights they are given when machines mediate the very considerations and actions we count on to uphold our values?

The point of this brief outline is to establish two premises: (1) that human tasks are capable of being performed by machine intelligence without any moral guide and (2) that ethical questions about data are directly relevant to those about machine intelligence. Machine intelligence is simply the results of putting data to use, often with no human oversight (e.g., statistical learning and

contextual adaptation). Much like the famous slogan "garbage in, garbage out," that warns of bad data leading to bad results, we must see that data ethics *is* machine ethics in an era of narrow AI. Within data ethics, questions such as, Should public online data be allowed for research? or What limitations should be placed on third-party sharing of personal data? are anchored by questions of machine intelligence such as, Should certain output classes be restricted without a user explicitly opting in to allow a statistical model to evaluate their data? or Should AI systems whose decisions could impact a human's life be trained on data licensed through a pass-through agreement with another service? We might wonder that if a massive dataset captured today carries a hidden statistical trend that men with no open political affiliations make for the longest-serving corporate employees, should we allow an algorithm trained on that dataset to filter job applications? Does it even matter if the features of "being male" and "having no political affiliation" are true predictors of long-term employment? I should hope our answer is, "no," if we were given the option to decide.

Perhaps, in a future where general AI is achieved, we will begin asking data-agnostic questions such as the nature of machine rights. For now, however, any inquiry that goes deep enough into data ethics will end up asking questions about machine capabilities and vice versa. One may even argue these are just two new languages in which we can cast age-old questions about fairness, openness, autonomy, and human rights. Going back to the advice of Zeynep Tufecki, if machines are now being granted the permission to make recommendations or decisions of moral importance, it is critical we agree on these morals, or at least know what moral perspectives are being loaded into the systems that shape our future. Otherwise, the narrowly-defined, superhuman systems we are designing may achieve their goals without regard to human norms, dignity, autonomy, or rights. Within the literature of ethics and risk, this problem was famously explored by Nick Bostrom, sometimes called The Paperclip Maximizer thought experiment. The gist is a superintelligent AI destroys the human race attempting to make the most paperclips it possibly can. Though, the idea of AI instrumantalizing humans [1] is rarely raised as a problem relevant to today's machine

[1] cf. the philosophical idea of "instrumental convergence"; primarily concerned with yet-to-be-seen superintelligence

learning practice.

A fantastic cultural example of this idea is presented in an episode of the popular sci-fi cartoon series, "Rick and Morty." In the season two episode "The Ricks Must be Crazy," Rick and Morty must go into a mini-universe that powers the broken battery of their space ship, Summer, Rick's granddaughter and Morty's sister, is left alone in the spaceship which is equipped with a super-intelligent computer. Rick gives the spaceship one command, "Keep Summer safe," before departing. While Rick and Morty are gone, the space ship ends up murdering one potential trespasser, paralyzing another, psychologically manipulating a police officer, and ultimately designing the conditions for a peace treaty on that planet, all in order to keep danger away from Summer who is inside the spaceship. As hilarious and absurd this representation of the problem is, it is substantively spot on. An AI that knows nothing besides how to optimize its model to achieve a narrowly defined objective, could very well act outside of any reasonable ethical or legal box to achieve its objective. Though the format of a TV show does not provide a framing or space such that serious debate could occur, it creates a common reference point for what it means for an AI to achieve goals using questionable means.

Before turning toward the role of narrative in this research program, let me solidify the ethical space this dissertation occupies. I am taking the stance that ethical questions about data and machine intelligence are fundamentally tied and that the implications of these ethical choices will involve us all given the growing pervasiveness of AI and data-driven systems. I am interested in pursuing two primary problems within this space. First, that ethical questions related to data and machine intelligence *must be* made explicit in order to not be tacitly defined and solidified via technical design implementation. An assumption I make here is that even expert engineers likely need support in understanding the consequences of their choices. Second, the moral choices we make *should not* be isolated to a small group of elite technicians and must be accessible for comprehension to the broadest audience possible in order to not alienate us of our fundamental rights. In Chapter 3 of this thesis, I will mount an argument based on human autonomy that drives home this point. For the purpose of this introductory section, it should suffice to state that

if we do not build a conversation where enough stakeholders can be involved, machines acting on our data may violate our norms and rights without us *prima facie* recognizing it. Not only is this unacceptable under a moral view that upholds respect of all persons, but it is an idea that could undermine our democracy if a populous were to be stripped of the ability to pursue their interests in the new social reality of Big Data and Machine Intelligence.

1.3 The Role of Narrative in Ethics Research

Establishing that there are serious ethical challenges for our society, and computer scientists in particular, is one matter. Why to explore narrative as a tool to meet these challenges is another. Restating from above, the goal is to find ways to ground ethical thinking about technical systems and maximize the number of individuals who understand the stakes. The notion of using narrative to communicate complex ideas is by no means new. There is a long history, particularly via science fiction (sci-fi) [59], of narrative being used as a tool for thinking about philosophical and ethical issues in technology [58, 307, 118, 211, 208].

Edwin Abbot's famous book, "Flatland," is an example going back to the late nineteenth century of a creative story about a society of circles, squares, triangles, and lines that helps readers understand the complex concept of *dimensionality*. The 2D characters encounter a 3D object beyond their comprehension. They are only able to perceive it as a circle that grows in size then shrinks again. Beyond providing utility for mathematical teaching, the story also critiques classist and sexist societies.

Weaving complex concepts into story form is may be a residual benefit or point of intrigue for some artists who have no didactic aims. A laundry list of sci-fi stories could be adumbrated (e.g., "Blade Runner," "Brazil," "2001: A Space Odyssey," "The Circle," "Super-sad True Love Story," "The Dispossessed"; to name a few) that are highly critical and thoughtful in their conceptualizing social and philosophical questions. These definitely set a precedent and show promise. Narrative or fictions that are useful for research, however, may need a more formal engagement structure. A mode of contextualizing into a literature, articulating an argument, or gesturing at real design

patterns can move pure art into a dialectic useful for research. This approach is adopted most notably by those who occupy the broad areas of speculative research and design. A list of terms under which you may find such work includes: "speculative design, critical design, design fiction, design futures, antidesign, radical design, interrogative design, design for debate, adversarial design, discursive design, futurescaping, and some design art" [118, p.11].

Designers Anthony Dunne and Fiona Raby attempt to summarize work done in this space in order to suggest that it elevate itself into a more central area of research and development. In their words:

"we need to move design upstream, beyond product, beyond technology, to the concept or research stage, and to develop speculative designs or "useful fictions," for facilitating debate. As designers, we need to shift from designing applications to designing implications by creating imaginary products or services that situate these new developments within everyday material culture. As the science fiction writer Frederick Pohl once remarked, a good writer does not think up only the automobile but also the traffic jam." [118, p.49]

It is here we find an articulation of what my work hopes to achieve in the space of technology ethics. Specifically, "useful fictions." Ways to get at the ineffable conundrums awaiting us in the future. Technical solutions and analytic debate that respond to tomorrow's problems. E.g., Pre-empting traffic jams. Before detailing my personal approach to applying narrative to ethics research, let's look at other approaches that create the context of this work.

1.3.1 Background of Narrative in HCI and Ethics Research

Through this thesis (specifically Chapters 4 and 5), there will be thorough discussion of the literature relevant to ethical problems in technology and creative approaches to research that use narrative. But before considering this literature as it closely relates to my research work done, let's step back and consider this field from a macro-vantage. Specifically, I want to look at how HCI researchers have been leveraging narrative and performance through "scenarios," "design fiction," "enactments," and "future studies."

Within the context of HCI, scenarios are widely recognized as a useful tool for envisioning how technology might function in actual organizational environments [82, 81]. Sometimes called "scenario-based design" [82], the idea is to construct brief stories that described how organizations or individuals will actually behave once a new technological system is in place and determine points of concern. Engineering software to meet the particular demands outlined in scenarios illuminates tangible goals for developers and expectations from stakeholders. The clarity of a scenario promotes the adoption of usable design [277] by allowing users to express their practical concerns in a common format that engineers can translate back into their work. It is in this vein that many HCI researchers have promoted scenario-based approaches to support participatory software development [340]. Again, as a participatory design tool scenarios are meant to allow stakeholders to elicit their requirements and engineers must correspondingly construct software patterns that can evolve as the requirements of a system design evolve. The scenario again acts as the common ground for user participation by allowing for the modification of scenarios to propagate into software modifications. "Vignettes" are another extension of scenario-based research in HCI, except vignettes often carry more context than the behavioral descriptions of a scenario. With this added human context, vignettes are useful for allowing people to reflect [327] on the human consequences of actions with technology or to study how engineers exert professional judgement [314]. However, this is not far flung from scenario-based research since again the vignette is posing a common object to translate between technological choices and human consequences. The summary point here being that scenarios and vignettes, both forms of narrative, are support tools within HCI research for relating software choices out to real-world behaviors and constructing spaces to negotiate and reflect on these choices.

Moving a bit more on target, HCI research has further taken up the use of narrative for a deeper reflective practice. "Design fiction," is a term first used by Bruce Sterling in 2005 [306], to describe his writing process in comparison to design thinking. It was not until 2009 that the term really came to capture its essence as a new idea for future-oriented research that straddles the line of fact and fiction [58, 307]. Sterling's evolution of the term, based on his own sci-fi writing,

came from the concern that sci-fi and design could learn more from one another if we allowed the imaginings of sci-fi writing to dwell more intentionally on technology that is closer to design reality than design fantasy. He writes,

"On occasion, sci-fi prognostications do become actual objects and services. Science fiction then promptly looks elsewhere. It shouldn't, but it does...However, when science fiction thinking opens itself to design thinking, larger problems appear...Many problems I once considered strictly literary are better understood as interaction-design issues." [307, p.21].

His call was simple: apply sci-fi to real design problems. Allow the fiction to motivate us outside of what the demands of commerce place on designers. Or as he puts it, "Rather than thinking outside the box–which was almost always a money box, quite frankly–we surely need better understanding of boxes. Maybe some new, more general, creative project could map the limits of the imaginable within the contemporary technosocial milieu." [307, p.24]. Unsure of what he was fully asking for, the insight was sound: unencumber our thinking about technological possibilities from the constraints of the market. Writing in ACM's publication *Interactions* his prompt was clearly to researchers. Asking them to see a new lens from which to study design, namely the fictional. And that approach has exploded into numerous projects. Two of the most notable practitioners of design fiction *research* are Julian Bleecker and Joseph Lindley.

In 2009, just two months before Bruce Sterling released his prompting essay, "Design Fiction," Bleecker published a similar article in *Near Future Laboratory* under the same title. The major difference, Bleecker was tracing a history where design fact and fiction are always influencing and playing off of one another. Bleecker's take

"As a principle, the science of facts and the science of fictions have their own distinctive characteristics which helps draw hard boundaries between the one and the other. But, in practice (which is what really matters when things are made), these two genres of science are quite tangled together. There are knots of intermingling ideas, aspirations and objects that blur any perceived boundaries and bind together these two kinds of science together. Engaging these knots, making the knots deliberately - this is the practice of design fiction." [58, p.25].

Unlike Sterling who is worried about wasting the insights fiction writers could give to designers (and vice versa), Bleecker wants to formalize the process he believes that is already there to elevate it to a central position in how we develop the near-future. More in the critical than the idealistic mode of thinking, Bleecker hopes to charter the future of design toward the good and away from the bad. How do we avoid that bad? By freeing ourselves of what we feel certain of through fiction. A clarifying passage of his position states,

"The kind of design I'm talking about is trying to determine with any certainty what will happen in the future. That's just silly. We're not interested in modeling behavior and saying with any sort of certainty or predictability what will happen. What design fiction is after is thinking through possible near futures based on a willfulness to create different worlds, perhaps more habitable, mindful of all the good things for which one might strive." [58, p.84].

Joseph Lindley, in response to these inviting words by Bleecker and Sterling, has taken on the challenge of really formalizing design fiction into research practice. His work on "anticipatory ethnography" [210] seeks to apply the rigor of ethnographic research to several dimensions of design fiction. If design fiction is meant to open up questions and pathways for thinking, anticipatory ethnographies are meant to document the aftermath of walking through that opening. Studying the process of creating a design fiction, the reactions from the audience, and the internal content of the fiction became the medium for much of Lindley's work [210]. Lindley further attempts to crystalize the practical purposes of design fiction as a research program after listening to the divergent meanings of the term floated through the research community muddy the water [207]. His insight is to see design fiction as falling into one of two framings, incidental or intentional. The claim is that the "incidental" fictions are those that catalyze discourse and create a space for debate by being studied. On the other hand, "intentional" fictions are deep inquiries where the design fiction and the research are co-created by and through one another.

Lindley's contribution has been an important one to separate out the process of making art and writing fiction that just so happens to raise questions about technology from the deliberate act of making fiction that targets subtle points and supports a research program. An example

of the latter project is the emerging HCI trend of "enactments" [248, 124]. Rather than asking questions about fictional worlds, enactments are design fictional scenarios where users actually become agents. William Odom set off "user enactments" as a hybrid research program that studied user attitudes and behaviors within designed fictional situations with technology. Their project starts by prototyping possible future scenarios where technology may put into question social norms then actualizing those scenarios and inviting subjects in to act them out [248]. The goal was to act out a scene up to a point then allow the participants to improvise their reactions to how the technology behaved. While Odom's program built toward moments of research inquiry, empirically examining what a user would do at the unknown juncture, Chris Elsden has iterated a step further designing whole social experiences with multiple stakeholders enacting the scene [124]. "Speculative Enactments" broadened the empirical study into the emergent behaviors made by an ensemble of participants working around designed, fictional artifacts in a scenario. For example, "Metadating" [125] studied participants who joined an experiential fiction where speed dating and revealing personal data were combined. People designed their own data-dating profiles to use in a "dating and future-oriented research event." The entire process was studied from how the individuals designed their profiles to how they interacted on the night of the event.

Stealing words from Elsden, a pithy way to describe these research programs might be: "Our role as design researchers was to create a set of circumstances where such speculation was was anchored in a familiar and relatable activity...with meaning for the participants beyond taking part in research." [124, pg.5389]. He later goes on to discuss this method as "consequential," where the participants are "co-constructing the fiction," and they "invite the study of experience." What I would like to point out about both Odom and Elsden's work is that they are able to take a future-orientation to their study of technology by designing fictions that their participants can easily adopt and understand. These studies then allow their investigators to take a stance on design and the future by not simply saying, "this is innovative, let's try it," but by arguing from evidence, "this is how people actually felt, let's try to avoid what was offensive and move toward what was well received." This kind of argument, about moving toward preferable futures, is part and parcel of

the final research method I want to discuss, "future studies."

Future studies is a conglomeration of methodologies all pointing toward one thing: understanding and steering the future of humanity. More robustly, "When thinking about the future is approached systematically, we can critically examine multiple potential futures, expand the set of externalities under consideration, and address both negative and positive forecasts of the future" [217, p.1629]. A common representation of future studies goals is represented by the taxonomy of futures graph that is used again and again by people in this space of research (See an example in 1.1). Using varying research methods from computer simulation to The Delphi Method (iterating on questions about the future using expert panelists) [217], future studies researchers draw from expert opinion, current events, quantitative analytics, and critical reflection to construct insights and recommendations about the future. Stuart Candy, now at Carnegie Mellon's School of Design, wrote a pioneering thesis on 'experiential scenarios' in which he details the pressing need to study the future and studies the the psychological impact experiential interventions have on people who go through his meticulously constructed workshops that invite people into possible futures [79]. Once again, we see the aim of studying the yet-to-be-seen, requires devised scenarios or fictions to support the objective of shared reflection about a common uncertainty. His work, much like a lot of future studies, is not oriented at technology in particular, but at the future broadly. However, his (and others') methods, as may be obvious by now, are being adopted and applied to technology-specific questions in HCI, design, and art.

Navigating through these threads of research, I'd like to now come back to highlighting a common theme that will be critical to my related exploration. Namely, that narratives are commonly used to (1) characterize and create possible realities for discussion and (2) construct common ground for multiple stakeholders to jointly experience. While much of the research discussed is not directly targeting ethics, it should be obvious by now that narrative may be an ideal candidate for digging into the unresolved legal and ethical questions posed by the emerging era of Machine Action and Big Data. Returning to the words of designers Dunne and Raby, "A speculative design proposal can also serve as a 'probe' for highlighting legal and ethical limits to existing systems"

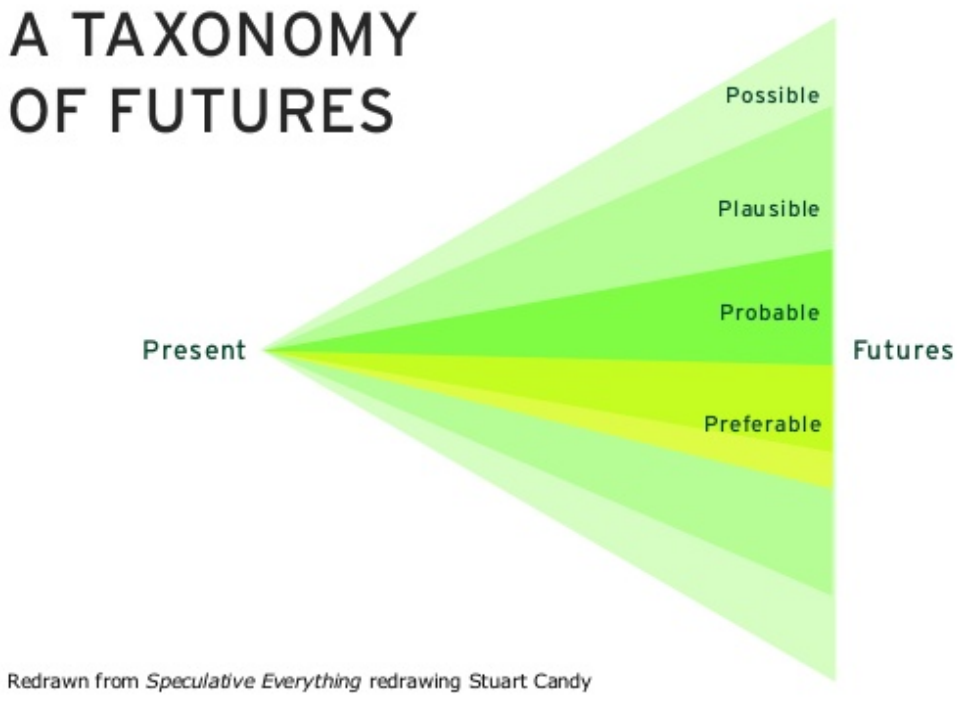


Figure 1.1: A graph originally drawn by Joseph Voros, redrawn here by Stuart Candy, that delimits the space of the future into the possible, plausible, probable, and preferable.

[118, p.54]. And this is exactly what my work hopes to do – create boundaries, targets, and templates for investigating these legal and ethical problems. Or, using my favorite phrase from Dunne and Raby, "This project could be described as a form of 'speculative ethics' - a tool for exploring notions of future good and future bad" [118, p.64].

This description of their practice, very accurately depicts my goals of leveraging fictional, future-oriented narratives to support ethics research. This thesis generously borrows from the above literature and hopes to expand upon it and specify practical applications, much like Lindley did for design fiction. Let's now move on to looking at my personal framing of this work and the investigative project of this thesis.

1.3.2 Speculative Ethics Using Templates, Targets, and Boundaries

This thesis showcases a variety of contexts, from policy framing to teaching ethics in the CS classroom to public dialogue, where narrative can support ethical inquiry. Targeting the complex issues arising from Big Data and novel Machine Intelligence, narrative is explored as a malleable tool with benefits within different contexts. Each chapter and section communicates a particular application into a space of problems where ethical thinking on these issues is important. It is possible to read the chapters as stand-alone essays where a problem context and methodology is developed. However, in this first framing chapter, I would like to give some insight into the overarching framing that structures my application of narrative.

Calling back to the language of Dunne and Raby, they frame speculative ethics as a "tool for exploring notions of future good and future bad." This idea resonates well with the way I conceive of narrative functioning in this research program. Another idea from the literature above that impelled my thinking was Lindley's work to separate out incidental from intentional design fiction. The dichotomy here is between narratives that amalgamate ideas that work to build the discursive space as opposed to narratives that are highly researched and curated to evoke particular thinking and enable research. For me, I think of speculative ethics, or the application of future-oriented narrative to ethical thinking, as falling into three species: templates, targets, and boundaries.

These species are not the narrative themselves, but the structure behind a socio-technical narrative meant to do ethical work in a research program. Let's consider the three in detail before going on to how they are used.

1.3.2.1 Templates

Templates are the source materials that creators (whether designers, writers, or researchers) use as a medium for thought. This could be a design artifact (such as a fake product), a computer tool (such as an adversarial machine learning system), or a piece of text (such as a written scenario or case study). Alone, templates may not fulfill the wholeness of a narrative, but begin to gesture toward a world. They are the objects that structure the creative boundary between fact and fiction from which to work. Often a template will be something that, alone, starts provoking intrigue, criticism, or ethical reasoning. In some cases, such as provocative scenarios, the template and the final narrative it goes toward will be very close in form. Perhaps the description of a plausible event simply needs a character and a bit of context to hone into the idea being explored. On the other hand, some templates, such as a product, may evoke a richer context, but be far from a narrative. In this case, a template and its narrative have a wide gulf between them as the template is too thin to evoke a rich socio-technical context. Templates are crucial toward the development of common ground between the creators and any audience that will ultimately interact with the narrative for analytic or experiential purposes. It is possible to develop templates directly into narratives for exploring problems without actually taking a position within the ethical space it occupies. Speculative Enactments, as discussed above, tend to be purely built from templates since the goal is to build a neutral space for co-creating the final narrative. And there design through research approach that creates the structure for further evaluation. Working from templates is great for an inquiry-driven approach (as Lindley may call, the incidental fiction) where the goal is merely to build a space for discourse and interrogation. However, if a creator wants to take a stance on a particular ethical issue, they may work these templates into an evolved form of a target or boundary.

1.3.2.2 Targets

A target is when the development of the narrative is directed toward a goal or a particular point in the future space of possibilities. Much like the role of narratives in scenario-based design, the narrative here is a way to represent the ideal case. In Lindley's words, these would definitely be intentional fictions. It should be noted, thinking about the narrative as a target does not mean that the creator is necessarily supporting this. Rather, they may be attempting to work out the complexity or impossibility of the ideal case really emerging. The role of the narrative is simple to express a trajectory where *some* stakeholder with *some* perspective might be trying to take a technology, and attempting to understand that world and aim. It is possible to go either way: the target may be an argument for a specific design or technological choice or an argument that is meant to question. The position matters less than the structure of the representation. Morphing templates into a target narrative is specifically the project of showing destinations. Places we might land.

1.3.2.3 Boundary

Another way to intentionally shape template material during the construction of a narrative is toward a boundary. Unlike targets which take us to a place, a specific landing point from which to look around, boundaries attempt to highlight where we may cross an ethical line. In making target narratives, we might tacitly cross many ethical boundaries that can later be considered from the new resting point. However, making a boundary narrative is the specific attempt to discover the point where the ethics get concerning. Much like case studies or court precedents, boundaries are where you shape your template materials for the purpose of getting to know that boundary condition so that others can see it. Boundary narratives are perhaps most useful when trying to make a legal or philosophical argument where the ethical concern becomes much more prominent within the context.

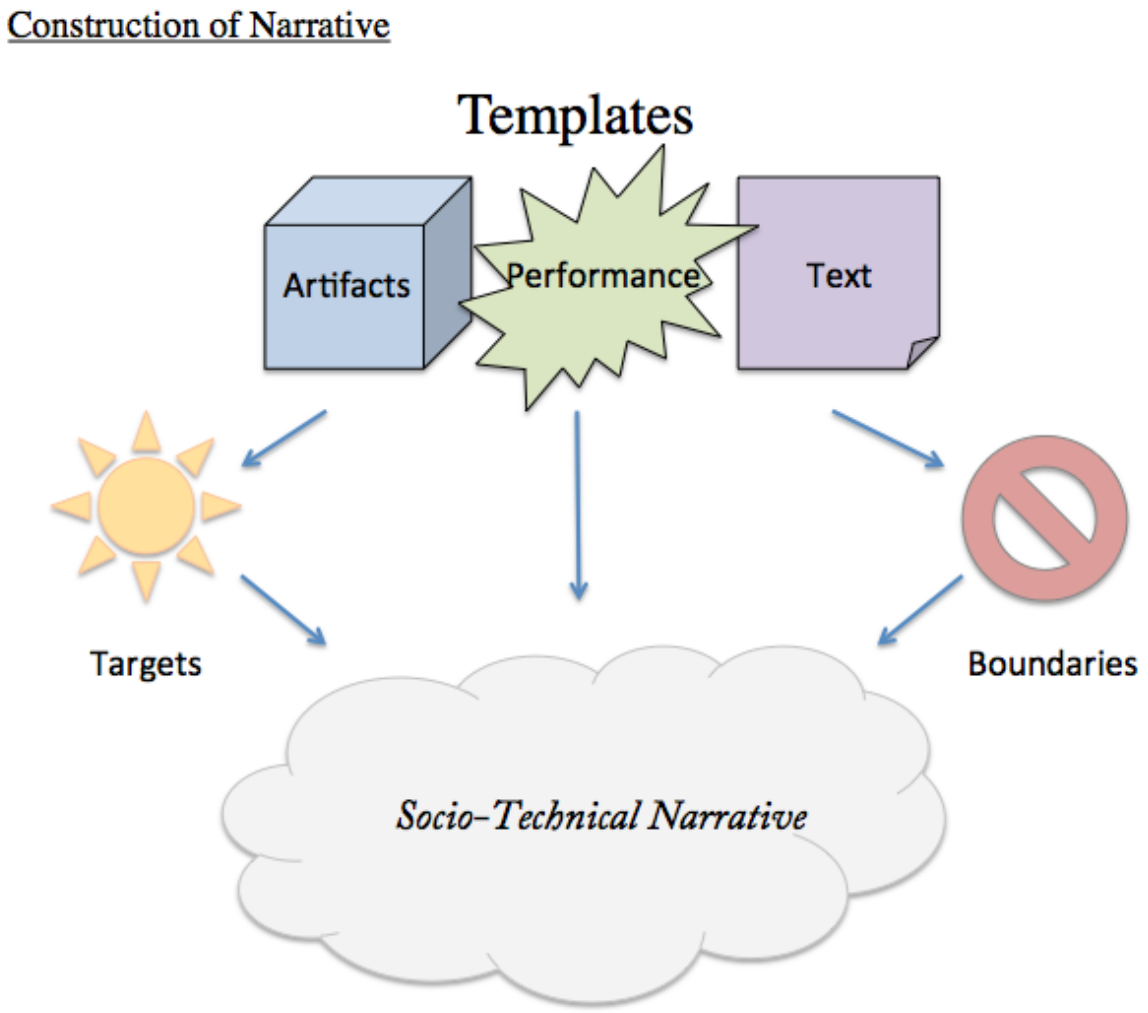


Figure 1.2: The network of templates, potentially shaped by targets or boundaries, that go into forming a socio-technical narrative.

1.3.2.4 Narrative Construction

To quickly ground these conceptual tools, let's consider a short example. See **??** for a visual guide. Let's say my starting template is the Amazon Echo Look. This is a real smart-camera product from Amazon that integrates with the Echo to allow you to take photos and videos of your daily fashion choices. This allows you to create "lookbooks," get selective clothing recommendations, and get AI-support to improve your desired style. By no means does a source template need to be a real object; though, often if the ethical inquiry is meant to be applied technological ethics rather than broad philosophical ethics, it will be. Some purely artistic texts and objects may get the mental gears turning toward a very real, applied issue, but it is easier to see the applied trajectory in this example by using a real product.

With this source template, I will want to begin considering plausible fictions regarding how the product might be used, ideally using different perspectives. In the world of the Echo Look there might be stakeholders such as a satisfied product developer, a worried software engineer, an insecure customer, and a rich, spoiled customer. Much like the HCI practice of constructing personas [263], the conversion of the template into the narrative will require human meanings and purposes. Often here, if the Echo Look aroused a particular ethical worry, say, imagining a very insecure person getting tricked into buying way too much or Amazon merging these photos with their newly acquired food purchasing data from Wholefoods, it is useful to ground this inclination in a human frame. This helps deepen the thought and test its reality. If it's almost impossible to imagine any person with the personality and psychology that grounds the problem, it's probably very farfetched. In whatever case, the goal now becomes to start fleshing out a world of behaviors, consequences, reactions, and emotions that is becomes your medium for thinking. This may be very easy if you know you are targeting (we'll get to converting this template to a target in a moment) a specific legal problem, or this may be the most challenging part of your research program, much like William Odom describes going through dozens of scenarios before deciding which ones were right for the User Enactments [248].

In the case of wanting to study the template as a space for inquiry, the goal would be to gather enough facts about possible worlds to then ground a site for contemplation. This could be your target narrative - no specific goal or problem - rather, a world where varying presentations, possibilities, and choices are at work. This would be a purely template-driven narrative construction where the evaluation may be to figure out what people's ethical qualms are through an interaction experiment or to create a playful space to see what behaviors may happen given a set of circumstances. Still the narrative is the common ground between participants and the creator. A template-driven example using the Echo Look might be to create a fake clothing store with changing rooms and vanity mirrors where several Echo Looks are giving people advice side-by-side or one-after-another. This idea may seem weird, but uncertain and the goal could be to purely build up a realistic experience of this to then study how people react to this space.

A different creator may take the idea of the Echo Look and be very worried about the longer-term use of all the data such a camera would have. Thus, the shaping of a target may be appropriate. That is, designing a narrative that takes us to the world where thousands or tens of thousands of people have Echo Looks and Amazon now has millions or billions of photos of people wearing their daily attire. A sample of this may be a story about a gritty, masculine man's evolution into a chic and stylish fashion guru. Perhaps in this story, his husband had been trying to get him to dress more fashionably for years and the Echo Look finally made it practical. This story might instead focus on the odd reactions of the people around him who are surprised by his transformation. In either case, putting out such a narrative for further evaluation would have the goal of a target. A provocation structured around a particular world containing the Echo Look from which people can ask questions and consider whether it is a place they would like to live or what the consequences of arriving in such a place are. I personally think of the TV show *Black Mirror* as a show that's all about narrative targets. They take single technologies and perturb them further and further out until we are in a world both familiar and not that can then be considered on a number of ethical and social dimensions.

Another creator may decide to explore a specific boundary area between psychological ma-

nipulation and advertising that could arise with the Echo Look. In this case, the goal would be to try and discover where that boundary occurs. What does it look like when we do not cross that boundary and the Echo Look is supporting acceptable, ethical marketing strategies? Then, in turn, what does it look like when we cross that boundary and the Echo Look is doing something more insidious and playing our emotions and daily ups-and-downs to exploit an emotional vulnerability? Once the boundary space is worked out, a narrative that is aimed at revealing the boundary would be created. Such an example may be a story of a woman who is struggling with anxiety and depression; now with the Echo Look in her life is glomming onto fashion as her coping mechanism. The narrative may suggest ways, such as facial expressions, language spoken to Alexa, etc where inputs may realistically lead a machine learning system into an optimization for recommendations that leverages emotional qualities. A contrasting human character may be added to see how a human picks up on and is sensitive to these emotions whereas the software underneath the Echo Look merely optimizes.

What these options–templates, targets, and boundaries–offer is a way to take a seemingly blank slate technology, such as the Echo Look, and place it in a space where values and ethics are in play. The Echo Look, whether acknowledged by its creators or not, does imply a problem framing and a solution. It's technical design–where it stores the photos, how it analyzes them–embed further values into the product. Stripping it down to a very pure engineering instrument–a camera that passes photos into a neural net–still cannot be freed of ethical discourse as someone must at least choose an objective or a set of input features and output classes. Often, none of these value judgments by engineers are made explicit and may remain tacit until a particular incident. And it is for this reason one may consider using these narrative strategies: to think about how the space of technical choices corresponds to the space of human consequences. To further elucidate and allow other people to share in pondering these challenges when the technical artifact itself might resist an immediate understanding of the associate problems. Thus the construction of the narrative takes a lot of cognitive work and, if being built into a strong research program, likely would require the support of a multi-disciplinary team of thinkers.

1.3.3 Interrogating the Narrative: Ethics Research and Dialogue In Practice

Elaborating my approach for creating the narrative is the first half, I now further look at the framing for evaluation, which is likely to be where discourse and research results would take place. In my work, which is ethics-centric, the goal is for the analysis of the narrative to drill through the many layers of thinking to arrive at a space for moral reasoning. For me, at the end of the day, this is where we can reach an ideal structure for discussion and negotiation. A space for the kind of democratic dialogue Dewey would uphold as fair and shared. I'll quickly borrow his words to motivate these analytic ends:

"It is not necessary that the many should have the knowledge and skills to carry on the needed investigations [advancing science and technology]; what is required is that they have the ability to judge the bearing of the knowledge supplied by others upon common concerns." [109].

In light of this, as will be seen in the layers of interrogation I use to evaluate socio-technical narratives, the analytic space is meant to overlap and connect both expert and common concerns. See **??** for a visual representation of the layers of analysis that can be applied to a narrative. Each layer represents a language or domain of inquiry that may unveil useful insights or spaces for discourse. Importantly, the moral layer is the lowest layer than anchors all other inquiries into a common space where we all have say. It is in this final layer where perhaps the most important type of inquiry emerges: negotiation. That is, the other layers such as *Sociological/HCI* or *Technical Feasibility* appeal to facts, expert knowledge, and careful analysis. On the other hand, moral reasoning is a normative type of reasoning that is subjective and political. It is also the kind of reasoning which should not be coerced or forced upon someone, but negotiated respectfully given other surrounding facts. Let's briefly consider these layers in turn.

The first layer of analysis that can be applied to the narrative most broadly is the literary evaluation. The fact it is the first layer in my graph is not to suggest it is the most shallow or least complex, rather that it is the layer that addresses the narrative for what it is as a world, a set of events, and an array of characters. This first layer looks at questions like, What happened? What

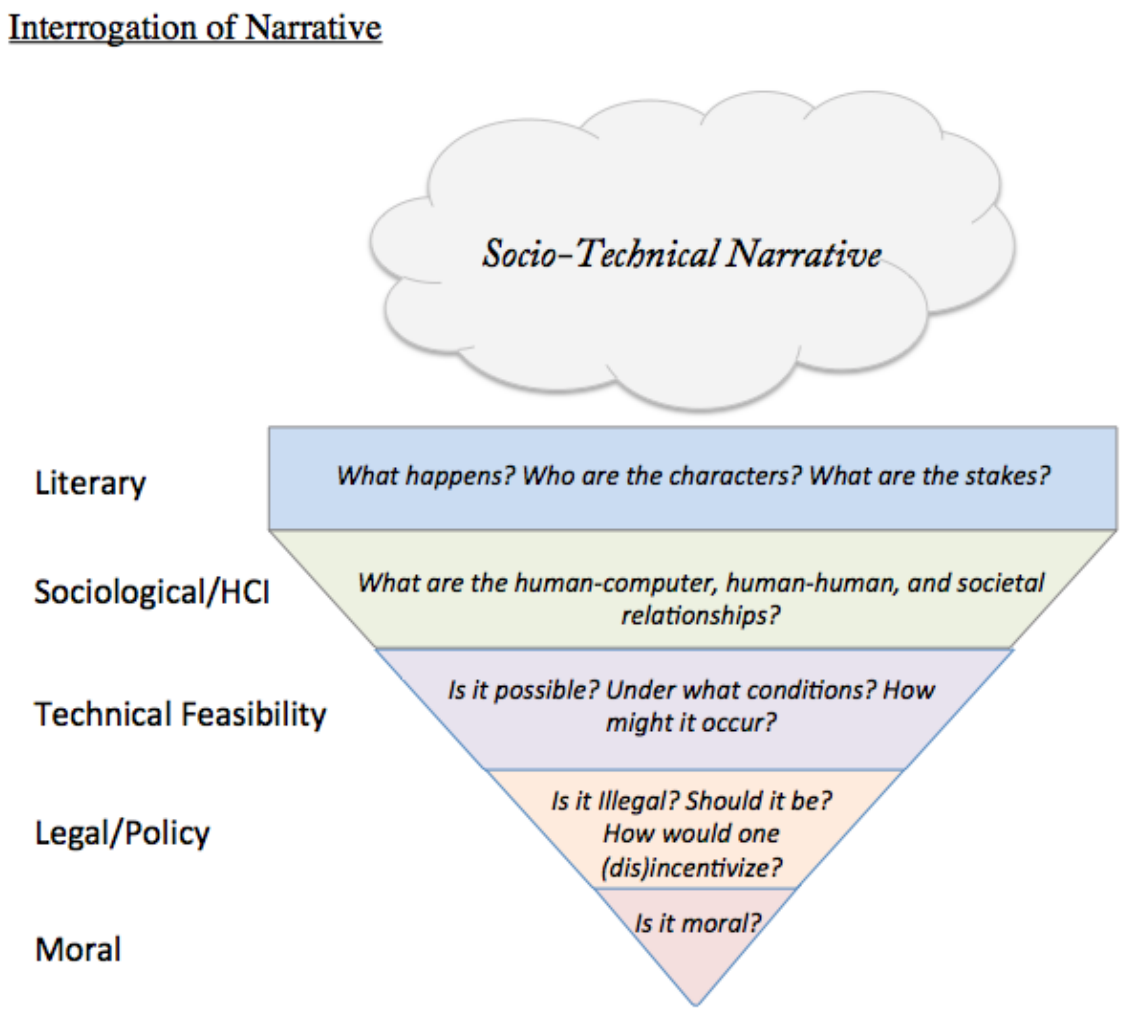


Figure 1.3: This graphic represents the evaluative questions that can be applied to a socio-technical narrative. Take note that the different layers may be best approached by different stakeholders; though the goal is that the lowest layer underpins them all and can be discovered within any expert or non-expert inquiry.

are the components of the world? What is it suggesting? Who was involved? What are the stakes? It is at this layer that we take seriously the components offered. This layer would also be applied to co-created narratives where a template was used for participants to explore and expand, a project where the final narrative was devised with the audience rather than offered up in its completed form. Following the event where the narrative achieved its final form, it would then be appropriate to start by breaking down what happened much like you would against a text, a performance, or a movie.

Next is the sociological layer, or for computer scientists, the HCI layer. This the space where we begin unraveling the socio-technical nexus. How did people interact with the technology? What human relationships did the technology mediate? What broad social dynamics formed in relation to the technology? For some projects, digging into this layer might be the ultimate goal. There might be ethical traces within this layer such as questioning whether the technology violated a social norm. Though ultimately questions about that social norm's standing, value, and moral relevance would penetrate deeper in this framework. It is important to note that taking ethical stances may not be the research goal and this layer generates plenty to chew on.

Lower we have the layer of technical feasibility. This is the reality probing exercise that is likely necessary for buy-in or sway given that if the narrative feels too off target or unlikely it's import may be ignored or blown off. The questions we might ask in this layer include: It the world or situation possible? Under what conditions would something like this happen? What series events would lead to this occurring? Are they likely? What factors would prevent or abet this? What technical choices are related to this world and series of events? I might add that this layer, I believe, is the most under-utilized by technicians. In today's machine learning environment, I cannot stop being reminded of sci-fi stories where computers are mediating our bureaucracies or promoting lazy thinking with life's most precious choices. And as I alluded to above, I believe we often allow the fantastical representations to push off the applicable substance. This is why this layer is very important. With some careful questioning, we may realize certain severe concerns are much nearer and more realistic than we would otherwise imagine.

Going down further, we find the legal/policy layer. This is where we might start doing more hardcore analytic and interpretative work. As was discussed, much of what's happening today in technology is not easily adapted to our laws. Regulatory institutes like the FTC have slowly adopted mechanisms or been granted abilities to take on challenges unique to technology; though, we still seem ill-equipped to handle this precipitous change. Thus, it is useful to ask questions about the legal and policy implications of the narrative: Is what happened illegal under current law? If it is currently legal, should it be illegal? What precedents are relevant? What kind of policy is incentivizing this? How might we imagine a way to disincentivize it? What stakeholders would fight to let this happen and which would fight to prevent it? It is only a matter a time until we see more policy, such as the EU's GDPRs, being wielded to control technology and its treatment of our citizens. And at any juncture there will always be emergent and difficult questions. Narrative analyses like these may improve our ability to be prescient in this planning.

Finally, at the bottom we hit the moral layer. This is in some ways the most complex and the most simple of the questions that can be asked. Complex because moral reasoning is going to be community-dependent, subjective, value-laden, and contentious. Simple, on the other hand, because it is something most people will easily adopt an intuition around. In fact, that is the goal of the narrative here. In the face of technical choices where we do not have intuitions the narratives support us in grounding our other ethical intuitions in the space of technological choices. In light of the narrative, we can ask: Was the person treated fairly? Would you want to live in such a world? Were the consequences harmful or violating? Would you be offended by such a circumstance? Is it just to let this happen? What is unique about these questions is that, unlike the rest, the inquiry is predicated on a second-person subject. This is because we are eliciting normative viewpoints. This last space of analysis is that of the negotiation. The rest leverage facts and expert opinions, which may still have some negotiation, but in this final space of morals, the discourse resolves to a common ethical negotiation where fact and expert takes may not matter so much.

Putting the whole framework together, we have **??**. At this point, it there should be sufficient clarity around the motivations, background context, and approach to the problem space of this

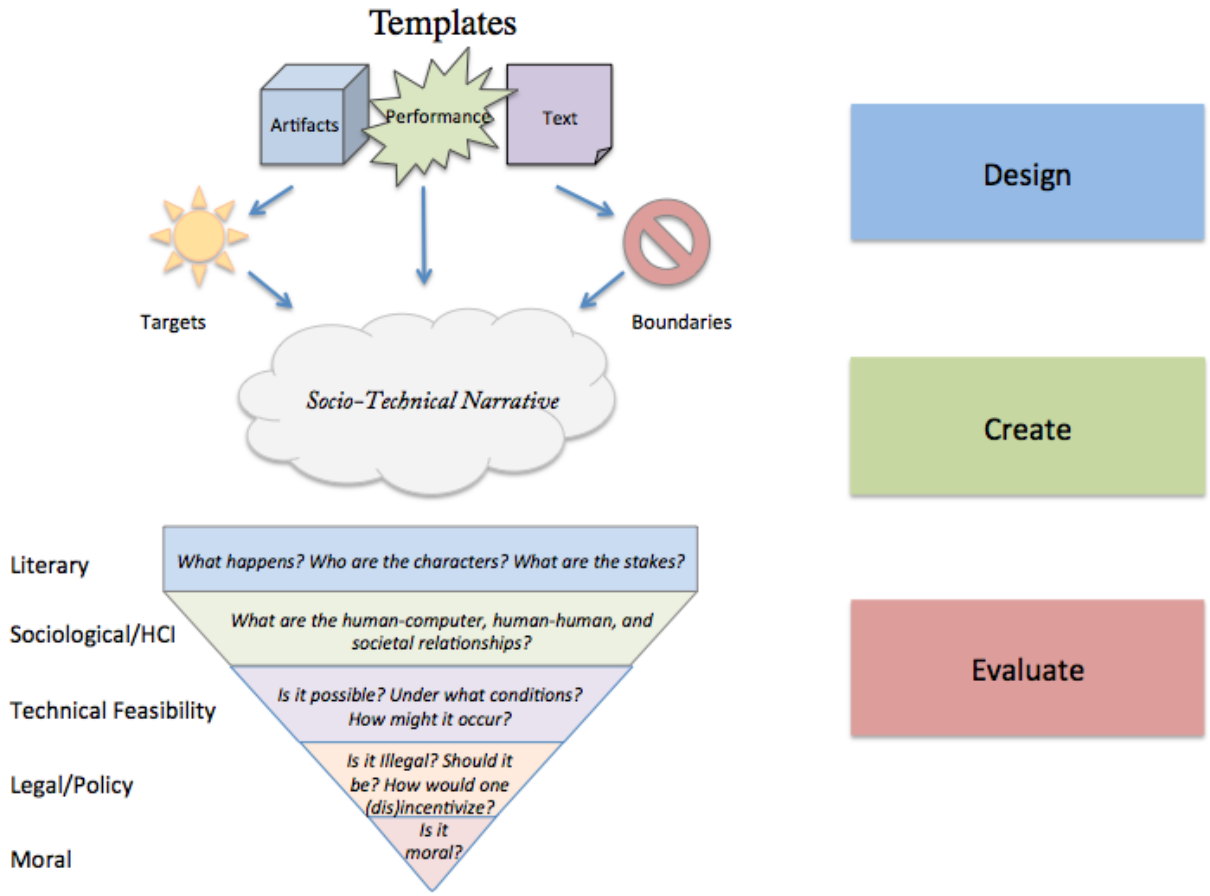


Figure 1.4: Here we have the broad representation of designing, creating, and evaluating a socio-technical narrative within my framing.

dissertation to move forward. I will now end this introductory chapter by walking through the rest of the dissertation and how it relates back to this research framing.

1.4 Outline of the Work Presented

Following this introduction, the work will be presented as follows. Chapter 2 is a deeper discussion of the ethical dilemmas we face in the era of Big Data and an argument for needing multi-stakeholder voices to make decisions going forward. This chapter is made up of two essays that were formative in establishing the motivation of this thesis. The first, written as my AREA exam toward this thesis, is a comprehensive literature review that goes through the ethical and legal work that builds the need for such a thesis. The second section is a paper published at KDD and presented at the Fair, Accountable, and Transparent Machine Learning (FATML) conference.[2] The argument is meant for an audience of machine learning practitioners who are trying to make fair systems. Though the topic is specific, the argument is general in that it establishes a need for multi-stakeholder dialogue to determine what is fair for a specific problem space. I include this in the second chapter of the thesis with the literature review as I believe it establishes relevant assumptions required to buy into the rest of the work.

Chapter 3 is building toward a regulatory framework for machine actions that is bolstered by a philosophical view of autonomy. It was written as a submission to the upcoming FAT* conference, which is a new ethics track for the Proceedings of Machine Learning journal.[3] This chapter has twofold importance. First, it establishes a moral grounding that underpins the remainder of the ethical work. Specifically, it sets a boundary condition by claiming that machine actions that undermine individual autonomy are unethical. Second, it shows the utility of narrative in the legal and policy context. Moving through the framework, I reach points where there are no definitive precedents for how unethical behavior may occur. Thus, narrative case studies are used to clarify these specific points that are plausible now or in the near-future.

[2] This essay was written with co-author, Micha Gorelick of Cloudera.
[3] The original essay was co-authored by Willie Costello from Stanford University.

Chapter 4 is geared toward promoting ethical thinking in computer science. Over the course of four sections, I show novel work that utilizes narrative for ethical provocation toward practitioners and as an educational tool in the CS classroom. The first section is a design fiction that will be published in the upcoming SIG-GROUP proceedings.[4] Straight from the design fiction playbook, the piece is written as API documentation for a fake product, Ad Empathy. The design fiction specifically addresses the problem of machine learning adapting to emotional vulnerability. The author statement establishes the real-world precedent for such an approach. The second section is a paper being published in SIG-CSE about a human-centered computing class I taught in the summer of 2017.[5] The paper presents novel activities used in the CS classroom to promote ethical thinking and training. Though the entirety of the paper is not specifically about narrative, the use of socio-technical narratives was a dimension of the educational intervention. Specifically, I had three authors write three short stories that considered future worlds extrapolating from single technologies. The students were to analyze these stories in light of class discussion. Further, the class was asked to write short stories to help them think about risk scenarios relevant to the technologies they were designing for the class project. The third section is another paper being published in SIG-CSE about the educational impacts of a theater project I produced about questions around data ethics.[6] This section establishes the value of the creative process for thinking and learning about ethical issues related to technology. The fourth and final section is a paper published in CVPR that leverages the experiments with a mix-methods approach to speculative analysis around the future of computer vision.[7] This essay attempts uses methods from future studies and risk perception research to try and analyze the likelihood and impact of a series of future risk scenarios involving computer vision technologies. After cataloguing and analyzing scenarios using the current CV literature, the paper leaves readers with two narrative case studies related to the scenarios that were found to be most concerning in the prior analysis. The goal was to help CV

[4] This work was co-authored with Casey Fiesler at University of Colorado.

[5] This paper was co-authored by Nathan Beard, Srinjita Bhaduri, Casey Fiesler, and Tom Yeh, all from the University of Colorado.

[6] This paper was co-authored by Jackie Cameron and Tom Yeh, both from the University of Colorado.

[7] This paper was co-authored with Tom Yeh from the University of Colorado.

practitioners think ethically about their field and offer up the narratives for further discussion and debate.

Chapter 5 considers how narrative can be used to research and work with the public. The first section is a paper submitted to CHI that compares the risk perceptions users and experts have about emerging technologies.[8] Micro-scenarios are used to describe future risks in terms of a possible harm. These risk scenarios are evaluated with a survey instrument derived from the risk perception literature. Participants were also asked to describe their worst-case scenarios they were imagining with respect to the scenarios that most concerned them. Narrative, thus, was used to both describe complex harms and elicit ethical opinions. Our goal was to learn whether experts and non-experts think differently about these situations. The second section is another paper submitted to CHI regarding what we learned from the public who attended an immersive theater production about data ethics.[9] The piece works to formalize the design space that combines technology and art into fixed and improvised structures for the purpose of engagement and research. We offer up heuristics for such projects, detail our project, and share the results we learned by surveying the audience and interviewing the cast and crew.

Chapter 6 is the conclusion of the dissertation. It briefly touches back on what was learned from the different research projects presented and how these findings relate to this initial framing. Finally, two smaller sections establish a guide for practitioners who may want to utilize my methods for education, artistic, or research purposes. The guide simply helps an interested person take this document and find what is relevant for them to spend time on given certain interests.

[8] This paper was co-authored with Tom Yeh and Casey Fiesler, both from the University of Colorado
[9] This paper was co-authored with Jackie Cameron and Tom Yeh, both from the University of Colorado.

Chapter 2

The Challenge Ahead: Why We Need New Ethical Thinking in Technology that Leverages Multiple Stakeholders

2.1 A Survey of Data Ethics: Problems New and Old

2.1.1 Prologue

This section acts as an extended literature review that details the many ambiguous and sticky questions in the realm of data ethics. For the dissertation, this section contextualizes and motivates my broader ethical inquiry in terms of a wide literature exploring issues from privacy to research ethics to open legal questions. The common thread is that each of these dilemmas is the direct result of Big Data's reshaping of science, institutions, and public life. Broken up into digestible chunks that focus on a family of issues, this section systematically highlights the current state of affairs within a burgeoning field here described as "data ethics." The questions discussed in this section will be revisited throughout the dissertation as they establish the domain of ethical issues for exploring with narrative.

2.1.2 Introduction

There should be nothing controversial about the suggestion that computer science and modern society are being reconfigured by new relationships with data. Machine learning, cloud computing, and network scalability are applied engineering areas undergoing accelerated expansion as data abundance and hardware innovations continue to open up possibilities and avail new insights. The

Fourth Paradigm of science has been named.[1] Any human interacting with an online system is now entangled in a constant process of being tracked, modeled, and solicited information.[2]

But these changes did not magically spawn out of a computational ether. Much like any technological development, applied engineering and research have been informing and responding to one another along a historical chain that brought about our Data Era. As far back as 1967, researchers at the RAND Corporation created the Relational Data File as a way to use a computer for "the logical analysis of large collections of factual data"[204]. Within the realm of HCI theory, there exists a rich history of discussion on topics such as contextual privacy, situated/desituated action, and grammars of action–the presence and capture of data being a substantive anchor in each conversation.

What has changed is that data-driven systems have recently been deployed like wildfire due to the growing availability of data-intensive infrastructure such as Apache's Hadoop and HDFS, Amazon EC2, and now Google's Tensor Flow. Within the past decade and a half innovations in computer virtualization, NoSQL databases, and GPU computing have pushed us toward new capacities for distributed, machine-intelligent systems. However, with any new technological and social arrangement must come a fresh set of risks, harms, and ethical norms. Some of these have been pre-empted; others we have only come to recognize in practice. For instance, I take it AOL did not realize how easy it would be to de-anonymize their users when they released a search query dataset back in 2006. On the issue, The New York Times quoted the executive director of the Electronic Privacy Information Center saying, "the unintended consequences of all that data being compiled, stored and cross-linked... are "a ticking privacy time bomb" [37].

It is this unforeseeable expansion in our notion of what's possible that has recently catalyzed

[1] The fourth paradigm was coined as a way to describe the shift in methodology happening in modern science. Beyond hypothesis-driven experimentation, scientific discovery can now be steered by data acquisition and analysis. That is, we can capture data without having any clue of what we are looking for to only later avail hidden information within the dataset. For a longer treatment of this see: Hey, T., Tansley, S., & Tolle, K. M. (2009). The fourth paradigm: data-intensive scientific discovery (Vol. 1). Redmond, WA: Microsoft research.

[2] As of 2014, the online advertising economy was at $120 billion. The fuel of this economy is personal data mined by online trackers, which is bought and sold by third-parties that may or may not have any interest in the original context. See more in: "Getting to Know You." The Economist, September 13, 2014. http://www.economist.com/news/special-report/21615871-everything-people-do-online-avidly-followed-advertisers-and-third-party.

researchers, engineers, and the government. Using the Association of Computing Machinery's (ACM) online digital library search engine, we find a growth in publications with the keyword "big data" going from 9 in the year 2011 to 223 in 2015. The Obama Administration began a National Big Data Research and Development Initiative in 2013. Writers at The Harvard Business Review have even called it early and determined the data scientist as the "sexiest job of the 21st century" [102].

Within the context of these new innovations, we must continue to ask ourselves, "what could go wrong?" Some of the discovered challenges fall into areas researchers, engineers, and legislators have dealt with for some time. The Belmont Report provides guidelines for IRB panels that oversee human subject research in our academic institutions. The US Government has designed regulations through acts like HIPAA and FERPA to protect individuals against the negligent treatment of sensitive information in health and educational contexts. In tandem, computer scientists have built their own standards such as RSA, SSL, and SHA-3 to ensure our ability to provide a secure online experience. What has only recently become clear is that our prior solutions to concerns such as privacy and consent cannot fix all the problems Big Data has exposed.

2.1.2.1 What are data ethics?

The long history of theorizing and debating about privacy and human rights in the context of computing falls under the larger category of "applied ethics." Applied ethicists are domain experts that project more abstract concepts such as rights, ownership, and duty into more specific subject matter. In what follows, a survey of applied ethical questions will be considered that are specific to matters concerning data and data-driven technologies. For our purposes Big Data will not be taken up as a particular quantity of data; rather, as a paradigmatic shift in the role data has in science, engineering systems, and everyday life. In this view, "data ethics" is an applied field encompassing the problems that arise alongside the accumulation of massive quantities of data and the corresponding changes to scientific and engineering practice. These are questions pertinent to engineers, HCI designers and theorists, lawyers, and users as we witness and become enmeshed in

this explosion of data-driven technologies. Domain experts have hardly had the time to reassess the relevant laws, social norms, and practitioner guidelines being challenged by Big Data. We now have risk assessment and predictive policing systems in place around the US that transpose our available data into a numeric score that courts can use to alter a defendant's treatment [27]. Trends like this put the heat on subject-matter experts and practitioners to quickly and soberly bring our attention to questions–new and old–regarding the ethical use of data in both engineering and society. Already, across the many domains touched by Big Data innovations, we are unearthing imminent ethical challenges in need of answers. In HCI and social science research, we are rediscovering ethical problems concerning collecting, storing, and sharing human-subject data. When using social media data, how does one get consent from a subject who does not even know they are being researched? When sharing data, how can we be sure a dataset is sufficiently anonymized given what information is now publicly available? Is informed consent relevant to strictly online research? Do industry data collectors need IRB approval to study their own data?

For computer scientists and engineers, we are constantly making choices with data that will shape other people. Is it fair to train a model using a particular dataset? Or will that dataset favor a particular group of people? When data mining, what metrics should be captured and how should they be classified? What risks come in deploying systems whose models are uninterpretable to the engineers who design them?

Lawyers and legislators further have to answer these questions for themselves in order to ensure proper protections and avenues of recourse are in place. Can an algorithm enact illegal bias or discrimination on protected classes of people? Who is liable when a computer model makes a decision that unjustly harms someone? Do users have any rights to own or control the data they produce?

And finally we have to ask fundamental, philosophical and social questions about what we want life to be like. Can we trust the most powerful technical actors to be benevolent with our data? Is there any danger in our relationship with proprietary search engines when looking for factual information? At what granularity and frequency do we care to be notified about changes

to privacy policies in the online systems we use?

2.1.2.2 Looking Across Domains

The following survey will frame and elaborate on the numerous issues being complicated and raised by data-driven technologies. The questions outlined above span a number of expert domains and express concerns that intersect with and depend on one another. However, in effort to cogently summarize such a broad topic, the remainder of this survey is structured into four distinguishing categories: 1) Privacy, 2) Discrimination, 3) Consent and User Attitudes, and 4) Algorithmic Impact.

The opening section on privacy will frame canonical questions about privacy in the context of Big Data. Having a longer history than other data-relevant topics, we'll see how older solutions to privacy questions have now proven to be insufficient and how those inadequacies bring light to new challenges. Establishing concepts such as "personally identifiable information," "anonymization," and "personalization" will in turn provide an important foundation for further elaborations.

From there, the focus will move to discrimination. Before the relevancy of discrimination to computing can be understood, an overview of the specific regulations that set the precedent for what legally counts as "discrimination" will be covered. The discussion will then move to contemporary engineering solutions that use machine learning techniques to train models for software applications. Any model that is trained requires data, thus we must look at how data may bias and affect the results obtained by our algorithms. Understanding bias, we can continue to see how often machine learning is a process overtly leveraged to discriminate and grapple with when discrimination is ethical or not.

Establishing a backdrop of foundational issues in privacy and discrimination will allow for a more lucid examination of questions directly concerning the user. The following section will introduce new problems surrounding consent involving terms of service, privacy policies, and human-subjects research. Grounding these concerns in present day controversies, attention will be paid to particular recent cases which have fomented critical discussion about the limitations of consent. In

light of these complicated cases, we'll also look at the attitudes and concerns understood from the perspective of the users coming out of contemporary HCI and social science research.

Moving up a level of abstraction, we will lastly encounter the topic of algorithmic impact. In this section, the nature of how algorithms modify and re-habituate our patterns of life will be discussed. After characterizing the vectors through which algorithms act on humanity, the remainder of the discussion will look at the broad consequences data-driven systems can have on society. Finally the essay will end by summarizing what's currently being done to mitigate these challenges and what future avenues we might pursue to improve our ethical situation.

2.1.3 Privacy: A Right and a Preference

Most notable among the list of issues raised by all the new data around us is privacy. By no means are privacy concerns new phenomena created by Big Data; rather, there is a long historical precedent governing the rights individuals have to privacy. In the legal sphere, the right to privacy interplays directly with a protection from harm. On the other hand, in HCI literature and our lives, privacy is a nebulous concept rife with normative judgments and cultural differences defining where we draw lines between the public and private spheres of life and in which contexts those lines hold. For instance, for one person the history of their love life may be a private concern, restricted to closest friends and never put on Facebook; whereas, another person may disclose their timeline of love publicly without any issue. The primary distinction which causes debate is between privacy as a right and as a preference. Specifically interpreting, "when has one's privacy rights been violated?" A naive interpretation of privacy rights versus preferences in today's online era would be to assume that there is a categorical divide between sensitive and insensitive data. Sensitive data are those special numbers and characters that we do not share to anyone but our most trusted confidants such as doctors, lawyers, and spouses. These should only be passed around by confidential means and those who obtain them should be heavily regulated. Outside of those special pieces of information, everything else is personal preference: share what you want about your life, but if it's online, presume its public.

Assuming we could conjure some standard for what's sensitive and what's not, we still face a number of challenges. First, what if enough insensitive data allows us to infer sensitive data? If I have access to your purchase history, might I be able to tease out attributes of your medical record? Or what about cases like the Netflix Prize Dataset where a dataset believed to be anonymized–ie, uniquely identifying data removed–was then de-anonymized using outside data sources, allowing for the recovery of sensitive information [238]? Both of these holes in our naive privacy view–the threats of inference and de-anonymization–are in fact real, tangible threats brought about by amassed data sources and new analysis techniques.

What this naive treatment hopefully shows is that basic assumptions made about our ability to protect privacy no longer hold. In the absence of these formerly-trusted techniques and protections, computer scientists and lawyers must search for new ideas. Let's begin unveiling these issues by first looking backwards at how privacy has been handled historically and slowly move forward into an understanding of how it is today's privacy scene has changed.

2.1.3.1 Legal Protections of Privacy

For lawyers, the provenance of information privacy law starts with an article, The Right to Privacy, by Samuel Warren and Louis Brandeis in 1890 [250]. The two lawyers argued that the rise of tabloid journalism brought about the need for privacy torts. That is, courts should allow plaintiffs to seek legal redress over the harms being done to victims unjustly exposed by tabloids. It took nearly seventy years, but finally the call for torts came in 1960 when William Prosser established four privacy torts that are still widely recognized in the US [262]. What these culminated in were legal precedents that one cannot intrude on someone's private affairs, publicly embarrass another over private information, conjure a false image about someone using publicity, or appropriate the identity of someone for external advantage. These torts operate on the notion of harm that can be done with information.

Had things stayed this way, computer scientists would have little concern since these violations presume some defendant intentionally enacted a privacy harm. Common privacy harms place no

burden on people who actually collect or store information; rather, the Prosser paradigm guides judicial determinations toward when someone calls attention to or actively exposes information for malicious purpose. Essentially, "no harm, no foul" is the doctrine when it comes to torts. Though as the government began to use computers for record keeping, a transition began from focusing on harm and to prevention.

The new world of digital storage erupted a novel set of questions about what information can be stored, for how long, where should it be kept, and who is authorized to use the computer. Over a decade of back-and-forth finally led to the "Fair Information Principles" (FIPS) and the legal ratification of those principles through the The Privacy Act of 1974. FIPS established many of the norms we find common today when collecting and storing personal information. Requirements such as notice and consent, individual's right of access, and individual's right of amendment along with corresponding enforcement and penalties became legally instilled [6]. A more fundamental change created in this policy was Congress embracing, "a wholly data-centric approach, the PII approach, to protecting privacy" [250]. Embedded in our regulatory approach to privacy was now an assumption that we could evaluate the risk associated with different data categories and precisely identify which fields in a database must be regulated.

Moving ahead in time, FIPS created a number of tactics to ensure these risky data entries would be appropriately handled. These tactics structured the foundation for now-commonplace anonymization approaches to sensitive data. Specifically, the 1996 enactment of the Health Insurance Portability and Accountability Act (HIPAA), which regulated the management of health records, designated a "de-identification of health information" standard [250]. Subsequent refinements such as converting birth dates to years and reducing zip codes to three digits were made standard, allowing doctors to share information without infringing on their patient's privacy. For academics this should all seem familiar since these standards were further adopted under the Family Educational Rights and Privacy Act (FERPA) and are now widespread across research as a risk mitigation tactic to storing personal data. We will later return to touch upon other aspects of FIPS, laying the groundwork for the privacy policies used by online platforms.

While small adjustments have been made such as the zip code and birth year changes mentioned above, the PII framework remains standing against the trampling stampede of Big Data innovations. As the acceleration of available data is growing and the locations where it's being collected more ubiquitous, the primary protection of our data amounts to specific columns of a CSV file being removed or obfuscated. Given this backdrop of data regulation, we can insert it into the context of HCI to grasp how these protections fit into the views espoused by the community of interaction researchers.

2.1.3.2 Contextual Privacy and HCI

Unlike lawyers who concern themselves with liability of risk when handling data, HCI researchers emphasize and explore varying dimensions of how humans might interact with data representations through computer systems. Thus, when it comes to privacy, major questions involve: What context does the user believe a particular action to be happening in? and How is that context communicated via the structure of the interaction? An example HCI concern involving privacy may be the use of automated notification systems. A pop-up notification on a screen may raise privacy violations for the person receiving the message, the person sending the message, or both. Importantly, the context matters since the pop-up is fully private if alone at a desk, but may cause professional damage when giving a public talk.

Paul Dourish describes the two primary HCI concerns of context as:

> The first is mutual relationship between physical form and activity; how we can design computationally-enhanced devices and how their form as much as their interactive ability affects likely patterns of action and interaction...The second concern...is how computation can be made sensitive and responsive to the setting in which it is harnessed and used. [114]

If our goal is to design systems that cooperate with a user's needs and enhances their abilities while performing some contextual action, we must be aware of the norms carried in a particular context of action. In terms of privacy, we are required to design interactions that safeguard users from harm (ie, we must follow FERPA) and should design with contextual norms and expectations

in mind (ie, we should hide notification content when a computer is in presentation mode). This means health monitoring, social network messaging, and email clients all face separate challenges to protecting privacy.

The problem of interpreting context is very challenging in and of itself. How are designers supposed to know exactly when and how a user will attempt to use their system? How can we adopt privacy standards that actually fit an uncertain context? Germane to these questions, Palen and Dourish have come up with a list of "boundaries" designers can use as guides in considering how a person might manage their privacy in a particular context [253]:

(1) Disclosure: what information may be disclosed through an action and under what circumstances?

(2) Identity: how is identity displayed and maintained for each party during a technology-mediated interaction?

(3) Temporality: how may the action be interpreted across the past, present, and future?

Leveraging the these considerations, HCI designers are meant to configure interactions that offer control and limit risk when confronted by these boundaries. Helen Nissenbaum applies similar points from a legal vantage. Arguing that universal privacy principles will cause too much disagreement, she ends up at a similar conclusion to Palen and Dourish: no principle is universal since cultures, opinions, and circumstances vary [245]. Or in the original HCI terms, boundaries are fluid across contexts. With this, she comes up with a tenet she calls "contextual integrity." Meaning privacy policies should seek, "compatibility with presiding norms of information appropriateness and distribution" [245]. Nissenbaum attempts to combine our legal history of privacy with the insights of HCI researchers, suggesting a reading of the law that equates privacy harm with violation of reasonable expectations of context. Implementing such a regime would require its own systematic (and debatable) categorization of contexts and agreed-upon standards. While these standards may bring an improvement to our current situation, by no means should we expect them to settle the issue given the wide disparities that exist culturally and individually [337].

Unfortunately for users of computer systems, no matter how much work goes forward pursuing contextual privacy by design, as it stands today, most information is immediately captured and prepared for uses not necessarily pertaining to the context of creation. The boundaries Palen and Dourish wish to interpret no longer take on knowable forms as disclosure and temporality remain uncertain at the moment of interaction. What we will go on to see is that neither PII protection and anonymization nor contextual privacy considerations are easily tractable in the world of Big Data.

2.1.3.3 Ubiquitous Data Capture

As things stand in 2016, data is being collected on consumers asynchronously, autonomously, and constantly. Of course, studying consumers is not a new phenomenon in business. Retailers have long shared information about who their customers are through subscribers lists and purchase histories [200]. What is different now is that data exchange is not isolated to conscious, isolated transactions. While typing this sentence, drafting this essay in a Google Document, Google is storing each revision made to the document. Most modern websites take part in behavioral tracking, social media networks convert our relationships into massive datasets, and mobile phones offer location-based services to add fine-grained geospatial details into the capture of online actions.

Consider a group of researchers' findings while testing for the kinds and quantities of data shared through popular Android and iOS mobile applications. Of 110 apps, 73% of Android apps shared personal information such as email addresses with third parties and 47% of iOS apps shared geo-coordinates with third parties [360]. Do all these asynchronous updates really correspond with the expectations of the context? Beyond the growth of data capture, the size of a payload has grown as well. Latanya Sweeney details the augmentation of person-specific information at a number of common locations such as the grocery store, during a hospital visit, and at birth . A particularly enlightening passage:

> ...a consumer in 1983 could purchase items from a supermarket and the only recorded evidence left behind were roughly an inventory debit and a record of

> the total amount purchased and the amount of tax paid. There was no knowledge necessarily of the identity of the consumer or of the consumer's personal habits and behaviors in terms of goods typically purchased and the times and days of the consumer's shopping experiences...Nowadays consumer transactions can be stored and analyzed, and by doing so, information about each consumer's lifestyle, behavior, beliefs and habits can usually be revealed. [310]

She calls the new policies technologists are adopting with data "collect more," "collect specifically," and "collect it if you can." Let's remind ourselves that these remarks by Sweeney are focused on explicit data transactions. Moments where an individual actually recognizes that an information exchange is occurring. Online machinations for behavior tracking abide by the same collection logic, yet often the person involved has no idea what kind of data is leaking out from them. A description of an online advertising interaction might look like:

> When a user navigates to a publisher who contracts with an ad network, the ad server simultaneously transmits an ad, looks-up the ad network's cookie in the user's browser, and logs certain information about that user's activity in a database. [40]

What this means is that online moments always have the potential to be more than they seem. Whether it's Facebook's "Like Button" lingering around on a page or an imperceivable pixel tag, the moment your browser parses an HTML file asynchronous callbacks are being sent to observing parties all across the internet.

Why is it that online services are so interested in squeezing every last bit of information out of us? Often it is their business model: they offer free content, they get paid by serving you ads. Advertisers only want to pay if the ads actually target a user likely to click, making behavioral profiling crucial to revenue generation. Other times the data is used to personalize your experience on the website. Collecting information about you may allow for better recommendations directly to you or make more robust inferences about people associated with you in your social graph. Most important to engineers, data is a powerful and necessary tool for training new models using machine learning or AI techniques. The more and better data you have, the higher likelihood your model will be useful (and profitable).

Establishing that data is being collected more frequently and in larger quantities does not immediately explain why these practices are harmful to privacy. If the data is handled correctly, why should we believe any harm is enacted? Personalization can improve computer interactions, so why not try to make better online experiences? What we now move on to discuss is how vast quantities of data makes de-anonymization and re-identification a much easier task and how personalization comes with a disclosure trade-off riddled with harmful potentials.

2.1.3.4 De-contextualized Action, Identification, and Personalization

Before we address these new privacy issues, let's recap. Earlier we discussed how the legal framework for privacy protection focuses on PII where specific information that could cause harm must be anonymized or left out before sharing. Further, we saw that in HCI privacy is considered an issue of context where we look to protect norms involved in particular uses of technology. The first thing to note pertains to context. Looking to HCI theorists Jonathan Grudin and Phil Agre, we find early warnings that contexts presented to users are slowly slipping away from their actions. As Grudin has discussed, the moment that information is put online, the context of "here and now" fades away [154]. The possibilities of how that information may be used or interpreted goes far beyond what can be imagined in that moment. Phil Agre, on the other hand, discusses how we reconstitute systems into particular "grammars of action" to support the desired data capture from the system [19]. He presciently warns us that actions will likely take on looser couplings with their built environments due to the separate concerns of data capture and human activity. Merging these two ideas with the above discussion about online data tracking, we see that the practices of Big Data has completed the arc of dislodging our data from context. Not only do our online actions go far beyond their original intentions as Grudin predicted, but online environments are now built to allow for data capture that is not congruent with the presented context. Bringing back Paul Dourish's first concern of context in HCI regarding the mutual relationship between form and activity, we should recognize that Agre's arguments had already pierced through the heart of his framework. The need for data capture is hegemonic on the logic of systems design. Respecting

norms of context has been relegated to superficial concern of appearance. Unless we somehow reel in the now-common practices of capture used by online platforms, a serious shadow of doubt is cast on the ability for contextual privacy to save us from violations possible in the Big Data Era.

Putting theoretical notions aside, let's take a step back to our legal protections. One may ask, "Even if droves of data are being collected, the really private information that could cause negative impact, that's secure, right?" Or, "Even if a bunch of decontextual data is piling up on servers somewhere, isn't it kept anonymous?" While the prima facie answer to these questions is *yes*, this does not quite salvage our privacy concerns.

We first must recognize that the anonymity criteria involves practices like deleting or obfuscating data entries that contain social security numbers, last names, home addresses, dates of birth, etc. What makes these entries important is that they allow for unique identification of a person to whom they relate. While it's very easy to identify someone if you already have their name or address, these are not the only pieces of information that uniquely identify a person. This fact leads us to a practice called re-identification attacks:

These attacks depend on a variety of methods: overlaying an anonymized dataset with a separate dataset that includes identifying information, looking for areas of overlap (commonly described as a linkage attack) or performing a sequence of queries on an anonymized dataset that allow the attacker to deduce that a specific person must be in the dataset because only one person has all of the queried attributes (differencing attack). [41]

Computer science researchers such as Arvind Narayanan have proven and formalized methods for de-anonymization of large datasets [238]. What this amounts to is any anonymized dataset having the potential to be de-anonymized by an attacker who has the right complementary dataset. And with the vast quantities of data anyone can get through the Facebook or Twitter APIs, it does not take long to legally and easily amass plenty of information to perform such an attack. In Broken Promises of Privacy, Paul Ohm argues that we have to transition out of a regulatory scheme that relies on PII as a central component [250]. Looking at the same evidence, he shows us that anonymization is no longer a way to protect citizens from privacy harms. A telling anecdote

comes from Barocas and Nissenbaum who say an engineer at Google claimed, "We don't want the name. The name is noise" [41]. And this claim should come as no surprise. Names are not all that unique; much less so than your browsing history.

Volumes of information can pragmatically be considered volumes of inferences. And it's the inferences that matter. Because it is very hard to predict what will later be inferred from any body of information, the constant accruing of online data only exacerbates privacy concerns. The power of inference goes beyond violations of privacy. A data set of Twitter Tweets, once conversations about daily life, may later become a tool for discovering rates of heart disease. It may appease researchers to say these findings are advancements to science and humanity, but that may not satisfy folks on the ground if insurance premiums go up in the locations determined to be at higher risk. As we will discuss at the end of this essay, the worrisome ability to de-anonymize datasets is leading to lots of new research around differential privacy, probabilistic programming, and client-side information storage. However, before moving on, there is one last privacy concern that big data uncovers even if you do not actually have the data: personalization.

The act of personalizing experience does not necessitate actually owning data on the specific user. Instead it could mean using contextual information to perform algorithmic methods like collaborative filtering or Bayesian classification such that you merely tailor content to the user. In many circumstances, personalization can lead to very positive experiences for end-users. Though, it is crucial to note that it also comes with certain risks. Recommendations often imply something about past behavior. Further, categorizing groups of people into clusters or groupings reveals information through association. These practices can often lead to moments of embarrassment and have the ability to compromise privacy. Take for example Facebook's Beacon advertising program. The program allowed activities such as purchasing a product or adding a product to a wish list to be published to users' friends' feeds. After plenty of negative media attention and a class action lawsuit, the program had to be shut down [320]. And these stories do not end here, Google's Buzz and Facebook's follow-up ad program Instant Personalization both had to be shut down due to privacy complaints. What this means for HCI designers is not that personalization cannot

happen, but that it's direly important to consider potential information disclosure and inference as a consequence of a personalized interaction.

What the above has shown is that our new data-driven society has complicated previously trusted notions of privacy. Both protected information and contextual integrity are challenged by Big Data practices. These concerns have many implications for the future design and implementation of computer systems. They also mean there is likely to be new legislation upcoming to try and minimize potential harm as these threats grow. We will discuss some of the more promising potentials at the end of this essay. For now, questions left dangling are how and when can we share data and retain privacy protection for our users' How might we impose limitations or make more transparent how data capture operates in certain contexts? What controls should a user have on future transmissions of their personal data?

2.1.4 Discrimination: Reorienting the Problem to the Machine

Much like a human who learns from experience, algorithms for learning may inherit unwanted bias coloring future decisions. One of the primary reasons for engineers to deploy data mining operations is to build statistical models for discrimination. A primary use of machine learning is to codify rationality into a model that can meaningfully distinguish between users and identify the features that make them statistically similar or different. Statistics is a science of distributions and data mining is a tactic to characterize a population distribution and make inferences based on prior information. So what happens when training data contains hidden bias based on gender, race, or class? Or what do we do when a model's recommendations differ across lines of identity?

To elucidate this point, let's roughly sketch a sample machine learning application. Assuming we are using some training data, and thus are performing a supervised learning task, the goal "is to learn a mapping from inputs x to outputs y given a labeled set of input-outputs D" [236]. D is our training set, the data we are using to "teach" our system about its task. Now for each input, x, we have a number of dimensions, or features, which are the attributes of the data (e.g., height, weight, eye color, age) we believe correspond to our output. For the learning procedure we will

need to define a cost function and a target variable. The target variable is the output we want our system to learn about and the cost function is how we evaluate how right or wrong the system is while training it. A possible application could be a system designed to identify good job candidates for a company. In this case, our input values would be gathered from an applicant's CV (e.g., age, university, prior employers, years of employment) and the system would output a score between 0 and 1 (i.e., 0 is a poor-fit candidate and 1 is a well-fit candidate). The learning process will involve us using past hiring decisions and training our model to correctly identify the past candidates that ended up being good employees. Seems like a great way to cull out a few good applications from the flood that applies for a desirable position.

In a field like engineering, dominated by white men, an immediate problem we may run into is that the model seems to only rank white men as well-fit candidates. This is a serious issue because in the US we have equal opportunity employment laws that protect certain classes of people from being discriminated on the basis of their identity. However, unlike the former days of an HR representative or a recruiter manually filing through applications, now an approximated mathematical function or a network of numerical arrays has computed a decision. And it is here we see how data has become a new issue for civil rights in our society. Machines are now in a position to enact the same kind of illegal and unethical decisions humans might make.

2.1.4.1 Legal Protections

Ahead of enumerating the different ways machines can discriminate, let's briefly cover some background on legal protections we have in the US. A summary of the laws applicable to machine discrimination was covered by the FTC in their recent report Big Data: A Tool for Inclusion or Exclusion (FTC 2016). They listed:

(1) Fair Credit Reporting Act (FCRA)

(2) Equal Opportunity Laws such as the Equal Credit Opportunity Act (ECOA), Title VII of the Civil Rights Act (Title VII), the Americans with Disabilities Act (ADA), the Age

Discrimination in Employment Act (ADEA), the Fair Housing Act (FHA), and the Genetic Information Nondiscrimination Act (GINA).

(3) Section 5 of the Federal Trade Commision Act (Section 5)

These laws were put in place due to a long history of civil rights movements that have attempted to equal the playing field in terms of opportunity and treatment in the United States. The FCRA applies to reporting agencies that compile and sell consumer reports. These reports are used as a basis for determinations regarding credit, employment, insurance, housing, etc. FCRA ensures reasonable procedures are in place that maximize accuracy of the reports and allow customers to access their own information and correct errors. Already two data aggregators Spokeo and Instant Checkmate have been successfully prosecuted under FCRA for discriminatory data practices that lack FCRA compliance [7].

Across equal opportunity laws, we find explicit protections for discrimination on the basis of race, color, sex or gender, religion, age, disability status, national origin, marital status, and genetic information. Using Title VII as an example, we find liability for discrimination applied under two primary strains: disparate treatment and disparate impact [42]. Disparate treatment is either proven intent to discriminate or evidence of formal disparate treatment of similarly situated people along lines of identity. Disparate impact applies to policies that prima facie appear neutral, but in practice have an adverse impact on a protected class of individuals. These two approaches to litigation under Title VII have historically proven to be difficult. Case histories have slowly moved the burden of proof onto the plaintiff. Explicit intent to discriminate must be shown or, in the case of proven discrimination, if the discriminatory policy is seen as a "business necessity" the plaintiff must show an alternative policy exists that achieves the same results sans discrimination. In light of big data systems, another layer of complexity has been added since discriminatory action is dislocated from any human and can be codified in an algorithm.

Finally, Section 5 is in place to prohibit unfair or deceptive acts or practices in or affecting commerce. Section 5 is a blanket regulation applicable across all market sectors and applies to

most companies. This is a consumer protection that prevents companies from making statements, designing ad campaigns, or omitting information that may mislead a consumer acting reasonably. More than protecting against predatory practices, a further protection offered by Section 5 is the sale of consumer data to customers that a company knows or has reason to believe will use the data for fraudulent purposes. The FTC has already prosecuted multiple companies that collect consumer data under this principle.

The scope of these protections amounts to a serious and encompassing regulatory system to protect citizens from unfair discrimination. Knowing these will help us investigate how data and common data practices have brought about new ethical challenges to discrimination.

2.1.4.2 Data as a Tool for Discrimination

ProPublica has recently begun a spotlight investigation under the title "Machine Bias." One of their central stories follows the the use of a risk assessment software called COMPAS [27]. The software is supposed to help determine whether a defendant is eligible for probation or treatment programs; however, they uncover several examples of judges using these scores in their sentencing decisions. In several cases, black defendants with lesser charges from prior offenses received higher risk scores than their white counterparts charged with the same crime. After hearing a few of these stories, the reporters did an analysis of judicial decisions in a county actively using the software, Broward County, Florida. They found that from the population of criminals who did not commit a subsequent offense 44.9% of African Americans were still labeled high risk in comparison to 23.5% among whites. And of those that were labeled low risk they found 28% of African Americans did re-offend against 47.7% of whites.

What makes this account even more frustrating is that COMPAS is a proprietary software meaning they do not have to disclose their methods. The potential for harm speaks for itself. There's an uninterpretable algorithm that uses undisclosed information from people's background to decide how likely they are to commit crimes and it just so happens it tends to more often deem people of color higher risk. Multiplying this concern is that it's very likely the engineers

who designed this system had no mal-intent. Given the inequality in prosecuting and incarcerating African American's in America, almost any training methodology using historical data is likely to find correlations between being colored and being a criminal. And this is why former US Attorney General Eric Holder [169], the FTC [7], and legal scholars [158] are all warning us of the potential harms data discrimination may have on our justice system. In Big Data's Disparate Impact, Solon Barocas and Andrew Selbst compile a list of engineering choices that may lead, in practice, to a discriminatory system. Though not exhaustive, this list is an important start for engineers or lawyers needing to audit data practices [42]:

- Training data that either over- or under-represents some class of individuals.
- Certain ways of labeling data can inadvertently or advertently cause your classification basis to be discriminatory.
- How you collect your data may skew toward certain populations by making assumptions about technological access or preferring particular behaviors or locations.
- Feature selection while determining the input variables for your system may insert bias by limiting what information the algorithm receives.
- Proxy variables that highly correspond to identity (e.g., school district with race) may end up approximating protected categories while not explicitly used.
- Masking can happen where someone who intends to discriminate can be removed from exposure by choosing a method that is likely to be bias.

Taking this list seriously, there are potential pitfalls throughout the entire design process for any machine learning system that will make decisions of substantive consequence. What makes this a particularly hard problem is the double-edged sword inherent in any attempt to regulate these actions. Imposing a liability on engineers would disincentivise exploring the field of machine learning and make systems design a quagmire given how difficult it is to interpret these risks. At the

same time, being lax about this problem could make hard-fought civil rights protections toothless whenever algorithms are involved in decision-making.

As intelligent systems are deployed more widely, the problem of identifying discrimination becomes more recalcitrant. Discrimination has already reared its ugly head in several other types of online interactions. Through a blackbox analysis with Google's ad system researchers at Carnegie Mellon found that being female makes it less likely to be served ads for high paying jobs [101]. In the same vein, Latanya Sweeney further found Google's AdSense to be much more likely to serve ads suggestive of arrest records and criminal history when performing searches on traditionally African American names [312]. Advertisements may not be deemed as important as financial, criminal, and employment decisions, but they often act as access points to services and carry along cultural information that is taken up by society.

At the moment, many other research efforts are beginning to assess the extent of discrimination in other important areas. Voter microtargeting has already burst onto the scene as a contentious issue [38, 166, 242, 271]. Already well established is that political campaigns are data machines that do everything in their power to get fine-grain information about their targeted demographics. Whether it's polling phone calls or online ads, parties and politicians work hard to change the minds of voters by any means necessary. An interesting project underway out of the University of Wisconsin-Madison is "Project Data." They install browser plugins that keeps track of political ads targeting you in hopes of getting a better understanding of how persuasion ads are being delivered, who pays for them, and who receives them.

Price discrimination has also put into question the nature of consumer rights [14]. Having a lot of information about your customers means you can make inferences about how much value a person places on an item at a given time. The demand for such knowledge is so high that Google even has a patent on a particular method for dynamic pricing. Perhaps useful for companies, this practice can be hugely unfair for consumers. Already price discrimination has been observed, but it is still difficult to say on what grounds the discrimination is occurring [228]. Since certain kinds of price discrimination are legal, such as adapting prices to current demands like commonly found

purchasing airline tickets, there is a significant burden on the consumer to determine when and if unfair discrimination is happening. Critically, we still are searching for what exactly one's rights are once the unfair discrimination is suspected.

This threat to basic civil and consumer rights has caused certain advocates like Kate Crawford to demand a right to procedural data due process [96]. As opposed to regulating on categories of personal data, due process would give a positive right to recourse given certain determinations made by algorithms. Any adjudicative process where Big Data plays a role in determining attributes or categories of the individual would be open to be contested in court. A right like this would be a huge step forward in ensuring individuals like the ones described in the beginning of the section at least had an opportunity to question the decision made by an algorithm. Big Data is on the advent of mediating our insurance premiums, mortgage rates, job and college decisions, and police interactions, yet we do not have defined modes of legal recourse for someone suspicious that they have been harmed by an algorithm. Issues like this seem destined for major legal determinations in our near future. As outlined above, there's a long history of government intervention on issues of discrimination. Therefore, it is simply a matter of time until a precedent concerning algorithms is set. Engineers joining in with the likes of Crawford to come up with best practices and regulatory suggestions will dampen the likelihood of that decision coming out of ignorance and thus having lasting negative consequences for everyone.

Summarizing questions to be answered: Should certain data be restricted from use within particular systems? What kind of transparency standards could allow citizens and juries to make sense of fair and unfair discrimination? Can data scientists come up with techniques to audit datasets and algorithms?

2.1.5 Online Research, Consent, and User Attitudes

Ethical principles for human subject research first became widely accepted following the Nuremberg War Crime Trials. The Nuremberg Code established standards to judge the work done by physicians and scientists on humans in concentration camps during World War II. Subsequently,

these rules would begin framing a long-term effort to protect participants involved in research. Similar efforts have been made to shape norms protecting consumers. Such is the case with Section 5 (discussed above), the Gramm-Leach-Bliley Act which obliges companies that offer financial products to explain their information-sharing practices, and the FTC's self-regulation principles for online behavioral advertising [3]. Central to both research ethics and consumer protections are the concepts of notice and consent. Specifically, "notice" pushes organizations who use or collect information to explain their practices so that consumers can "consent" by making a meaningful choice of whether or not to participate (ie, opt-in/out protocols).

The norms around notice and consent are manifest in the form of IRB protocol for researchers and Terms of Service agreements and Privacy Policies for commercial actors. However, as Big Data practices continue to enter into more aspects of life and society these standards have come under scrutiny. As research using online data continues to grow, questions around consent have started to surface. What does it mean to do research on human subjects who do not know they are being researched? How do we effectively communicate opt-in and opt-out choices to massive online subject pools? Is it even fair to use data that was never meant for research purposes?

In the corporate world there are fewer standards which has led to public outrage regarding some recent work proving that private companies are experimenting on their customers without notice. Moreover, Privacy Policies and Terms of Service are often nothing more than click-through agreements completed without the consumer batting an eyelash. And so the question emerges, "As consumer data collection becomes a bigger part of online life, what does it look like to provide meaningful consent to consumers?" Perhaps even more concerning is, how do we actually relay the information in a way that users understand the consequences of their consent?

This section will be dedicated to understanding the regulations and practices currently at work regarding consent, how Big Data practices are challenging those norms, and what we have learned so far about how users actually feel about these systems.

2.1.5.1 The Belmont Report and the Problem of Online Research

In the aftermath of horrible injustices committed by researchers on vulnerable communities such as in the Tuskegee Syphilis Study, necessary provisions were created to grant protections to participants of research studies. The Belmont Report [280] initially laid out the guiding ethical principles for human-subject research and Common Rule legislation in 1991 explicitly regulated them through Title 45 of the Code of Federal Regulations [325]. The results of these government interventions are crystalised in the form of the Institutional Review Board (IRB). These boards, in place within academic institutions across the country, operate as oversight committees to review, approve, and require modifications to research activities applicable under the Common Rule. The ethical principles stated in the Belmont Report and applied by IRB committees are "respect for persons," "beneficence," and "justice." Important to our discussion is how the IRB has traditionally interpreted the "respect for person" tenant through means of informed consent.

Before the advent of large-scale online data collection tactics, informed consent was fairly straight forward. Not to say disagreements and difficulties did not emerge, but mostly informed consent was about being sure to clearly assess and state the possible risks of taking part in a study and then communicating those to your subjects. The result is usually a lengthy form that each subject reads and signs giving the researcher legal rights to perform the intervention and keep the resulting data. In light of new availability of public data, however, this traditional method of consent is proving to be insufficient.

Nowadays more people are participating in online spaces such as social media networks which offer relatively easy ways to collect large amounts of data through APIs. This novel infrastructure is accelerating possibilities of online research through social media, online forums, trace ethnographies, text mining, and activity monitoring [332]. Meanwhile, as online research efforts are ramping up, concerns of the new risks coming to light from Big Data are beginning to stimulate fresh ethical conversations about how to adapt new principles [332, 288, 363] . Informed consent with online research has largely gone out the window due to practical difficulties and a lacking imagination

of possible risks. That is, most online research is deemed exempt by the IRB. When collecting information from thousands or even millions of Twitter accounts, how does one meaningfully communicate research intentions? Unlike highly controlled experiments in lab settings, online subjects enter into research on an asymmetric footing. They are there to socialize, learn, play games, and generally do the activities of their daily lives while researchers often only enter after the fact with separate interests unknown to the participants. The first headline case of online research gone wrong involved a 2008 study called Tastes, Ties, and Time. Within days of the project's first data release, which included information scraped from 1700 Facebook profiles, properties of the dataset were allowing people to be identified uniquely [363]. As days went by more aspects of the dataset were de-anonymized until finally the researchers had to retract the data.

Pertinent to the prior discussion about the robustness of anonymization, what these researchers learned was the extreme difficulty in removing all unique identifiers from a dataset. Since the time of this mishap, work has been done to get a handle on where the Belmont Report fails to engage with extant problems in online research. Jessica Vitak [332] studies the ethical norms held by people working in different disciplines using online research. She found lacking shared principles across research communities and a general belief from academics that they were held to higher standards than industry researchers. Further, she has pointed out that we are in dire need for new creative means of transparency and encouraged continued ethical conversations between online researchers.

While this academic discussion continues, industry has begun to take advantage of their large datasets. In 2014, Facebook released their results on the now-famous Contagion Study [195]. Modifying users' social feeds, where users receive updates of recent activities across their friend network, Facebook researchers wanted to see if positive and negative emotions were contagious. By selectively showing users messages with more positive or negative sentiment, they studied subsequent posts to see if the content invoked similar emotions in the user. With no IRB board, Facebook deemed the study ethical on their own terms. And, given that they own the data, there was little that could have been done to stop them. After the public erupted in anger, OKCupid came out

in solidarity with Facebook admitting that they too do experiments on their users [279]. It turns out, behind the scenes, the popular dating site was conducting a number of experiments including suggesting people as matches who actually were not (based on their algorithm).

Straying from weighing in on whether the studies were actually harmful, the point is clear that our networked interactions are being toyed with by industry researchers who have no real oversight. The sheer acknowledgement of these studies forces us to reevaluate the place of the IRB and regulatory systems applicable to human-subject research. Now that online software products are sites of monitored and stored behavioral action, do there need to be new provisions to determine who is required for regulatory compliance? For the time being, notice and consent are the bastions of protection and fair choice. The current solution to this problem depends on those lengthy terms of service agreements we all have to click through before registering for any online platform. And this brings us to our next topic: notice and consent enacted through online terms of service and privacy policies.

2.1.5.2 Privacy Policies, Terms of Services, and What Users Actually Think

Whether in implicit or explicit terms, the moment one takes part in a networked interaction that interaction is governed by the policies of the hosting web domain or software owner. We call these delineations Terms of Service (TOS) and Privacy Policies. Practices vary around where and when explicit consent is required, but it is common to require a user's conscious consent during registration for an online service or at the moment of installation for software. TOS outline the specific rights a user has in relation to the service and the special details of how the service is meant to operate. Sometimes privacy policies are included in a TOS, but they are usually separate declarations of the entity's information practices. The primary distinction here being the handling of information (privacy policy) versus the general rights of use (TOS).

Returning to a legal framework from our privacy deliberations, FIPS also sets practice guidelines for what must be provided to the user to fulfill notice and consent. Namely, FIPS demands users be, "given notice, that is to say informed who is collecting, what is being collected, how

information is being used and shared, and whether information collection is voluntary or required" [41]. There is no mention of specific rights that must be offered nor standardized language that must be upheld. Currently we are in the wild west of online contracts. Most regulators and practitioners keep their fingers crossed for a day where plain language and easy-to-understand policies are standardized.

Critical to this hope, Barocas and Nissenbaum discuss The Transparency Paradox. They claim that simplicity and clarity is a trade-off with fidelity. Each piece of software or web platform has its own complexity that requires certain edge cases and caveats. This makes it extremely hard to settle on a single set of terms or practices that universally apply. Pushing us further, Barocas and Nissenbaum argue that the very idea of notice and consent is infinitely difficult to track in the online world. Developers are constantly trying out new features, business deals emerge, contracts expire, and all the while, the terms the user signed do not even apply to the third-party services shared with by the original provider. It's as if I tell you a secret, then have you sign a contract saying you can only tell Billy and Cindy today, since I know they are trustworthy, but don't bother with whom Billy and Cindy might tell nor if you decide to tell other people a few weeks from now. The chain of who touches someone's data and what they do with it is indeterminate and unpredictable.

Adding complexity to the issue, even if we could come up with some legal standards, there is mounting evidence that users don't understand these contracts, and when they do, they don't actually want what they are being offered. A number of researchers have shown that copyright implications of TOS are misunderstood [135], privacy policies are too complicated for common users to read and digest [213, 175], and many users do not even want the forms of personalization offered using their data [323, 29]. One study found that only 11% respondents could understand a text description of an opt-out cookie, a common mode of control offered to users against tracking [223]. The same study found only 20% of participants wanting the "benefits" of targeted advertising. Google researchers submitted a study to CHI showing that users do not believe they are receiving extra benefits by allowing their data to be shared to third parties [55]. A mere 23% of nearly a

thousand participants in their study believed they received benefit if the first-party company uses their data and a dismal 6% thought it benefited them if the data was shared. Another researcher at Pomona College [25] found 59% of respondents to a survey about web tracking believed websites collect too much data. The same respondents widely agreed upon support for stronger do-not-track options (92%), a requirement to delete personal data (96%), and real-time notifications of tracking taking place (95%).

The contrasting facts of a) users not understanding what they are consenting to and b) when surveyed they often do not want what's being offered, is concerning. Our only real resolution of this seeming paradox is to say users should quit using the same services. While this pragmatic mindset functions logically, some researchers argue people feel powerless given the social pressures and lacking alternatives [25]. The felt powerlessness seems realistic given that many privacy advocates have pushed for better do-not-track policies with minimal success. Of many legislative attempts only California Assembly Bill AB 370 has gone through, which merely requires websites and services to disclose how they respond to a do not track signal (and only in the State of California).

It's unclear whether notice and consent can hold on as our ethical solution to online research and contractual user agreements. As has been suggested several times in this essay, data's use value is often unknown at the time of collection. This complicates the very idea of a one-time verification of consent when its ideas and commitments are ephemeral. For HCI researchers, this leads to a major challenge upcoming of how to improve user knowledge of what actions mean in networked systems. Legal experts, on the other hand, will need to ultimately decide what rights a person has while online. With these solutions still in limbo two central questions to focus on are: "What does it mean for someone to be informed enough to give meaningful consent?" and "How do we manage consent over longer time periods?"

2.1.6 Algorithmic Impact

Perhaps the most common way which a human is affected by data is through interactions with an algorithm. In a previous section, the problem of discrimination was taken up as a potential

harm that can result from the determinations of an algorithm. Discrimination is a specific issue that fits snugly into a more comprehensive legal framework built long before our current techno-social lives. What we now take up is the multifaceted ways in which algorithms restructure aspects of our life and have long-term consequences on our society. Though the broad topic of algorithms could stretch wide, we will focus in specifically on the category of algorithms which are directly constructed from and adapted to data. Questions about when to use a greedy algorithm or dynamic programming solution are not of concern. Instead, we may wonder whether there should be concern about the hidden and proprietary nature of Google's search algorithm? Or how might the use of PageRank as opposed to EdgeRank influence epistemology? One way to describe the algorithms of interest would be "public relevance algorithms" [147]. These are computations "whose inputs are composed of our personal and collective activities, expressions, and preferences." Crucially, we must refocus our thinking from algorithms as technical means with a fixed purpose to society in favor of artifacts with their own morality and ideology. A simple example is to consider a search algorithm that preferentially indexes websites based on the number of known links to a page as opposed to the quantity of string matches within the page. The first assumes that if there are more links to a page, it is more important; whereas the second assumes that the page that uses your exact phrasing is likely to be more important to you. Widespread use of either will have tangible consequences to what information people access and therefore to what people know and believe. What this means is that we should not simply consider *the search engine* as an isolated phenomenon that solves a single category of problems. Rather, we should consider aspects such as how a particular choice of what data to use, what relationships between data are privileged, and what transparency mechanisms are offered to determine the ethical standing of a particular algorithm. In this section, brief ethical considerations will be introduced to highlight the many locations data-driven algorithms are making their marks on us as individuals and communities and on our society as a whole.

2.1.6.1 Calculated Publics

A term beginning to be used by researchers such as Kate Crawford and Tarleton Gillespie is "calculated publics." As opposed to "networked publics" which are communities assembled by new mediums of networked interaction such as online forums and video games, "calculated publics" are those communities implied by groupings and suggestions made by algorithms [147]. Kate Crawford [94] meditates on the meaning of searching for a book and being suggested a series of other books under the heading: "Customers who bought this also bought..." There already is a semblance of a book-buying public that is constructed from something like The New York Times Best Sellers List. But unlike a bestsellers list, whose members we intuitively understand, what is the relationship between being derived between the people buying books on Amazon? Should we be reading the other books those like us read? Is anyone paying to make sure their book is connected to the book just bought? Perhaps more impactful is a category of profile visibility users are able to choose from on Facebook: friends of friends. This choice raises questions of trust within your social network. Are my friends' friends likely to be good people? What kind of people do I accept as a friend on Facebook and do those people use the same discretion as me? Gillespie [147] brings up a more concerning set of human relationships implied by Google's search algorithm when he searches the term "she invented" and Google asks "Did you mean he invented?" While Google did not hardcode that into their system, their learning algorithm was able to learn our culture's misogynist tendencies and is now imposing them back on us in a search recommendation. Similarly, advertising constructs its own communities through grouping people by taste and interest. As people wake up every morning and go online to read the news, many people originating on different websites will be suggested the same article or presented with the same click-bait headline. Using a payment website like Venmo will automatically begin showing you people's recent financial transactions who they believe are related to you. Certain algorithms like collaborative filters or Markov chains inherently associate people by saying *other people who have attributes similar to you liked this* or *other people in the same position as you did this next.* Again, there is nothing wrong with these algorithms, but

as our communities, friendships, and public groupings are shaped by them, we may ask questions about their appropriateness. Should algorithms that associate people use filters that ignore negative stereotypes? If we start accepting recommendations as objective, may we become vulnerable to unwelcome advertising embedded within? What choice should people have to disassociate from an algorithmic grouping?

Knowledge and the Structure of Information

Networked access to information has largely been celebrated as a triumph of the internet. Websites such as Wikipedia have removed major financial and physical barriers to knowledge being disseminated equally. While these advances should be seen in high regard, we must also ask whether there are risks associated with converting our investigative and reasoning skills into digital means. Moreover, as data stacks up, how are we meant to interpret all of it? How can data be manipulated to make arguments seem valid without actually being so?

The first example to consider comes from a recent study by the American Institute for Behavioral Research in Technology. They were interested in finding out to what extent biasing search results could change the opinions people formed. Further, they wanted to know if people would even notice that they were interacting with a biased search engine. After studying people in different global locations during real election cycles, they found across the board 20% of undecided voters were influenced by biased results with certain demographics being even more vulnerable [127]. Perhaps even more concerning, the presence of bias was entirely undetectable by their subjects. They have coined the term "search engine manipulation effect" (SEME) to describe the phenomenon. Considering realities such as the creation of The Groundwork, a Google-backed data operative working for the Clinton Campaign, results like these should warrant some pause. Moreover, this should engender concerns about the consequences of the well-known "filter bubble" or "search bubble" problem. Now that most search engines look at your past searches, location, and profile information to determine what you are likely to want, it is becoming easier to only see the arguments and opinions you already hold.

Another recent happening that places into question how we trust data is the Volkswagon data

scandal. The EPA found that Volkswagen was manipulating data during emissions tests. By using physical characteristics of the car to automatically determine when the car was undergoing a test, the car could activate different emissions controls than what would be found on the road. Thus a basic data collection practice used for environmental protection has proven to be falsifiable so long as it's possible to figure out when it's happening. With auditing practices as a widely accepted method for oversight, this scandal should bring into question how it is we can validly audit any system using algorithms to tailor itself to environmental variables. As if making sense of scientific findings was not hard enough already, now with the advent of Big Data, the possibility of p-hacking has become of heightened concerned [31]. In traditional scientific investigation, one tests the validity of their hypothesis by conducting an experiment and seeing whether or not their intervention had measurable effect. A commonly used metric to determine if the treatment group was significantly different than the control group is a p-value. However, now with an easy accumulation of massive data sets, it has become possible to simply search for correlations that have significance then slap an explanation on it. Inversely, if a scientist has a theory they want to prove, they can collect lots of data and simply choose the subsection of it that contains the correlation best fit to their theory. Some questions raised by these trends are: How do we assess the validity of a data source when it's far too large or complex for a human to manually go through? Should there be public standards on what can go into a search engine? Should we regulate relationships between online information providers and institutions such as political campaigns and research groups?

2.1.6.2 Inclusivity and Exclusivity

As already discussed, algorithms require some discerning choices in regard to what data matters to it and what does not. Once put into the context of different systems, we see that these choices often amount to ideological commitments. For instance, in the context of predictive policing, should my political views be relevant to a risk assessment of my criminal likelihood? Choices like this must be made in almost all data-driven systems. Take for instance a controversy that arose with Amazon a few years ago. Attempting to be family friendly, Amazon does not index adult

books in their sales rankings and recommendations. This seemingly understandable practice led to quite the political statement when Amazon decided to categorize all LGBT-related books as adult and overnight removed their presence in sales rankings and recommendations [147]. More recently, there was a conspiracy theory floating around that Google was manipulating their autocomplete feature to hide scandals by presidential candidate Hillary Clinton. However, Google responded that there was no scandal, but they simply do not allow offensive or disparaging autocompletions. While the conspiracy theory is wrong, it is interesting to reflect on how the choice to remove certain terms effectively shifted the appearance of objectivity. It avails that Google had determined they have license to decide for us and our families which lexical relationships are fair, kind, just, etc.

"Shadow bodies" has become a useful term to express the relationship our physical bodies has to the person described by our data traces [147]. Personalized features and anticipatory mechanisms embedded in systems cause software providers to encourage us to share as much as possible. Salient to this practice is what information is deemed relevant to a human profile and what is not. Keeping track of someone's search queries and email tendencies may tell a very different story than knowing their actual favorite books and hobbies. The former may have traces of the latter, but could also be misleading. These choices of "what matters" are being made autonomously such that the consequences may be out of even the developers' control. This was the case when Google recently found their image tagging software labeling African Americans as "gorillas." Unable to determine what features their image AI system had codified as important, they simply removed the tag from the system.

Concerns to reflect on due to these trends are: Should users be able to decide what categories of information are collected on them and for what purposes? How transparent do content providers need to be when making choices about what to show and what not to show? When algorithmic services hurt or embarrass someone, should anyone be held accountable?

2.1.6.3 Data as Power

In her provocative essay, *Can Algorithms Be Agnostic* [94] Kate Crawford raises a higher-order question: "How does the device or system generate a means for authority and/or power?" She claims that disagreement and dialectics are the heart of a healthy democracy. The concern is that as we replace systems of human deliberation with systems of algorithmic determination, will we lose critical sites for political struggle? The advent of data-driven systems may prove very useful when lost in a new city, but could they be doing us a disservice when they choose not to show us information on the protest happening 30 minutes from our doorstep? In relationship to these questions, ethicist Lucas Introna has argued for what he calls "disclosive ethics" [171]. Essentially he asks us to keep account of the moral implications of pragmatic and technical decisions, "at the level of code, algorithms, and the like–through to social practices, and ultimately to the production of particular social orders, rather than others." His argument starts with a recognition that algorithms do not always disclose their presence or intentions to us. He distinguishes between transparent (e.g., a garage door opener) and opaque technologies (e.g., embedded face recognition software) to expound the need to keep track of technological deployments in our environment that otherwise may be difficult to see. Disclosive ethics essentially amounts to having genealogies of technological choices and their consequences such that we can maintain a transparent view of "how we got here" when we run into a particular problem.

Crucial to this essay's discussion of data and the above ethical dilemmas is the recognition that data is power. Choosing to give a certain company or institution your data is an act of power changing hands. Data is the fodder of science, it's the brick and mortar of machine learning models, and it has the ability to come back into our lives in complex and undetectable ways. In the mode of disclosive ethics, we can see from the history of companies who own or control a particular avenue of data capture, those companies will have a competitive advantage in the future market of related algorithms and systems. Taking this line of thinking a step further, the visions of the designers and engineers who take ownership over our data have the power to shape the future of our society.

Someone may argue that anyone with a sharp vision and some skills may be able to do this. The rebuttal to this is to consider a problem like speech recognition: no matter how innovative an interface or algorithm is, it will be an uphill battle to compete with companies like Apple, Amazon, or Google who already own all the best training data.

It is in this spirit of reconsidering our rights and the rights of data, this essay moves ahead to its conclusion. The final section will discuss some of the promising avenues of research and engineering that are attempting to tame the chimeric animal that is Big Data.

2.1.7 The Future of (Data) Ethics

The story painted by the above deliberations may make the future of Big Data appear grim. Portents that privacy as we know it may be dead, the reinvention of identity based discrimination, and the adaptive swarm of internet algorithms out for your attention–taken in isolation, look bleak. On the brighter side, these same methods may lead to major advances in medical diagnostics, easier and fairer exposure for small businesses and artists, and a reduction in bureaucratic overhead on time-consuming, perfunctory tasks. The future remains unknown. What is taken to be the case in this essay is that there is a partnered relationship between knowledge and control. The more we understand what is happening in this rapidly changing technological landscape, the better potential we can harness it for human betterment rather than wake up to a world for which no one feels they signed up. As engineers and designers conjure up new visions of how our world could be, those possibilities must be evaluated with careful, critical eyes. We may soon come across the day where a machine decides whether a bomb drops, a car swerves into oncoming traffic, or whether you should get that line of credit to start the business of your dreams. As the interrelationship between human and machines deepen, it's crucial we forewarn ourselves of what risks are being borne by our users and society. Perhaps even more so when the innards of those machine "minds" are built out of bits of information we emit through our daily lives. Looking at the issue through this lens, the machines are mere approximations of us. The questions then become which approximation, whose machine, and for which purposes.

Before we conclude this wide examination of concerns raised by the Big Data Society, let's briefly see what progress has already been made in thinking through these issues.

2.1.7.1 Privacy

As we explore ways to recover privacy protections being attenuated by new accessibility and capabilities of Big Data, several engineers have already pitched promising solutions. One solution to the privacy conundrum comes under the name of "differential privacy." Proponents frame the problem as such: some trusted party wants to hold onto a dataset filled with sensitive information, but want to allow others to access statistical and global information about the data. Providing real aggregated statistics may lead to de-anonymization attacks and thus, it comes into question whether the data should be shared or stored at all. Differential privacy claims there is no need to use, or even have, the raw data in order to provide statistical information to third-parties. The proposal is to use select randomization algorithms that take a dataset as input and outputs a new dataset that retains the same statistical properties but differs in terms of single elements [119]. In this way datasets can retain much of their practicality while not openly exposing information about the people being represented.

For engineers who want to create personalized systems, but do not want to bear the risks of owning abundances of user data, client-side storage may be an option. The idea is to store all the relevant data on the user's system and only call it into RAM when a personalized feature requires it [320]. This may not solve problems for systems where the goal is to train a model, but it could be used in the case of features that simply want to evaluate past queries or locations while not storing them externally. It also would allow the user to control the persistence of the information by having the choice to delete the memory associated with the application. In practice, engineers could still capture, but not persist data externally, allowing them to feed data to their models for training, but then only be able to evaluate their models upon user interaction.

A separate standard to consider is Latanya Sweeney's [311] k-anonymity. Her technique is to remove information until there are k entries that cannot be distinguished. In the case of a medical

database, you could imagine if all entries regarding heart disease were reduced to patient's sex and primary symptom, there would be a lot of overlap across the dataset. Sweeney adopts this idea as a method for analyzing when enough removal or obfuscation has occurred. What number k is may change with the size of the dataset, the context of the data release, or other factors.

2.1.7.2 Consent

Deriving a guarantee for informed consent may be the most challenging issue of all since it is predicated on the idea of a universal standard of knowledge for all users. However, a few recent research efforts have shown interesting results. One idea was to allow privacy policies to be socially annotated so that users who are more expert could share knowledge and elucidate concerns to privacy policies [36]. In a first exploration of the idea, researchers found users reporting a higher level of comfort after reading through annotated policies; though not necessarily a higher degree of competency with the ideas afterwards. Though still in its early-stages, we could imagine a space opening up where users can share information and debate relevant questions publicly in relation to privacy policies. Having these annotations could also be helpful in the event of a future court injunction that requires some historical evidence or in the event of a privacy policy changing and users' wanting to compare.

Another group out of Carnegie Mellon has introduced the idea of the "privacy nutrition label" [183]. The essence of the idea is to transplant the standards the FDA put on food labels into the realm of privacy policies. Users would see standard evaluations of areas that may be concerning such as copyright and licenses, third-party sharing, and anonymity. While the idea is attractive, we as a society still have not determined the requisite language and standards to make it useful.

2.1.7.3 Legislation and Oversight

Slightly ahead of America, the EU has already begun implementing laws to protect users' information online. The first landmark attempt has been the infamous "Right to be Forgotten." What this amounts to is a positive right for users to request that data be removed from the

public internet and unlinked from search results. Requests are handled with discretion around how problematic the information is, whether removing it would adequately address the concern, and how much time has passed. By no means is this law perfect given this can both help and hurt: it may be bad that powerful people can effectively erase the internet's memory of their misdeeds. Regardless, this is a nice experiment to see whether such an effort mitigates certain harms done to people by slanderous or embarrassing information being taken offline.

Separately, the EU has very recently passed legislation slated to go into effect in 2018, which gives citizens a "right to explanation" [151]. This idea resembles Kate Crawford's priorly-discussed notion of "data due diligence." The law both places limitations on how autonomous or algorithmic systems can make decisions that "significantly affect" users and gives users the right to request an explanation of how the decision was made. It's not clear yet how this will work in practice, but the idea definitely progresses us in terms of accountability under the threat of unjust discrimination by algorithms.

Most recently, a consortium of researchers from Harvard, MIT, and the University of Zurich have released an article outlining Elements of a New Ethical Framework for Big Data Research [329]. Many of their ideas have been touched upon in various places above, but it's worth offering an overview since it may be the most comprehensive account out there of a way forward. Possibilities the authors propose are:

- Universal Coverage: put all research, both industry and academic, under the same oversight regulation and create new participant-led boards to handle the overload this would put on IRBs.
- Conceptual Clarity: enforce specific language to standardize terms such as "privacy," "security," "sensitivity," etc and declare a formal method for revising these terms as technology changes.
- Risk-Benefit Assessments: advise or enforce internal systematic risk assessment and train outside parties who can conduct reviews and create guidelines for review processes.

- Standardize Procedural and Technological Solutions: adopt a list of approved standards for technical and algorithmic implementations, privacy policies, etc so researchers have mandates on what's acceptable practice.

- Tailored Oversight: journals and communities can establish required reviews that must be passed before publishing or granting. Tiered access to data can be mandated based on the sensitivity of the particular context.

- Multistakeholder Processes: Setup boards, consortiums, and other conglomerations of people in expert domains and put them together with people who represent privacy and research interests to set recommendations or perform oversight.

Naturally none of these solutions offer a panacea for how to fix our ethical challenges with data. In fact, many of them would be quite hard to implement and oversee without risks to over-complication or nepotistic determinations. However, the suggestions create a starting point for people who are seeking solutions to these problems and may provide useful for future legislators responsible for making major decisions.

2.1.7.4 Where We Go Now

A lot of ground has been covered throughout this survey. Hopefully it provides both macro- and micro-viewpoints into the challenges our new Big Data Society has brought us. As a closing remark, I would like to offer up a few questions that should be seen as central to future research and work in this area.

- What rights should an individual have to the data they produce?

- What would a meaningful opt-out choice look like?

- Should there be an auditable trail of who data has been shared with?

- Under what circumstances should data records be deleted?

- Are their categories of decisions unfit for algorithmic aid?
- Should weaponized AI ever be allowed?
- Should humans ever be entirely out of the loop for decisions of significant impact?
- In areas with historical injustice, is it possible to train unbiased algorithms?
- How do we communicate risks to users and the public?
- How do we communicate about data and technology to retain the Public?s ability to make meaningful choices with technology?
- How do we communicate the long-term implications of a user sharing their data?
- Is there a fair way to inform the user about uncertainty?
- What does a fair risk analysis look like for data-driven algorithms and systems?
- Should risk assessments be public for all consumer-related technologies?
- How do we think about risks that may be long term or not yet applicable?
- How do we bring users into the process of assessment?

Considering the nature of ethical discourse, we should never expect that these questions will be perfectly answered. The best we can hope for is an establishment of agreeable norms that create a relationship of trust between an individual offering their data and the entity taking ownership. With the growing ubiquity and commonality of data-driven systems, it seems likely these "data ethics" will slowly become indiscernible from the more general category of applied ethics. Hopefully with the growing need will come a growing interest by researchers, programmers, designers, and lawyers.

2.2 The Authority of "Fair" in Machine Learning

2.2.1 Prologue

This short piece is an argument toward the engineering community attempting to focus the budding area of Fair Machine Learning. Fair machine learning is a sub-field of machine learning where engineers are working to engage with issues of justice and discrimination through quantitative means by training algorithms that can operate under certain constraints. Tacking very important issues, this field often forgets to expand its thinking broader to not only as *how* to make machine learning fair, but *what* is fair in different contexts. Thus, the paper argues for an approach to fairness that requires multi-stakeholder voices in order to have the authority to use a fairness construct within a particular problem space. This focused argument is also a motivational one for this thesis. Namely, this paper introduces a second motivation for the paper: beyond ethical inquiry, we must engage multiple stakeholders in our discussion. The argument here can easily be generalized to other normative concepts besides fairness and provides a strong basis to justify my desire to use narrative as a tool for broad engagement in ethical issues .

2.2.2 Introduction

The recent boom of machine learning (ML) applications has just as quickly given rise to a slew of critics pointing out the harmful capabilities of these systems. In particular, concerns of bias and discrimination are being debated as ML systems for natural language processing [62], judicial sentencing [27], target advertising [312, 101], image classification [5], and facial recognition [74, 319] have all proven their ability to inherit bias and create disparate treatment across groups. Responding to these findings is a body of work that attempts to import considerations of "fairness" into our ML approaches [20, 84, 172, 178].

In this paper, we argue for expanding and deepening our approach to "fairness" in ML practice. Drawing from philosophy and ethics, we offer up a normative account of fairness where "fair" is a property that is both communally derived and context dependent. Using this definition,

we highlight three categorical framings through which one may inquire about the "fairness" of an ML system: fairness of a system, fairness of an approach, and fairness of a result. We justify these different framings and then move on to overview contemporary approaches to fairness in the ML literature. We argue the position that the literature has thus far focused on the problem of disparate treatment without much attention to other framings. Making this salient allows us to consider the importance of critically examining "whose version of fair" we privilege in ML moving forward and argue that taking a stance on fairness must be understood as invoking an authority that could be more or less legitimate. We conclude by offering possible pathways forward for the community to broaden its approach to fair ML.

2.2.3 The Construction of "Fairness"

There was once a time when fairness or "the good" was believed to be a static and essential property that could be derived from divine [258] or rational principle [106]. Determining what was right or fair was a privilege vested in certain authorities and the results were absolute, allowing for no disagreement. This traditional view has changed as modern philosophy and ethics has revised our typological ways of thinking into a normative framing where concepts such as "fair" or "just" are no longer static; rather, they are developed relative to particular communities (who) and contexts (when/how). This is why for modern philosophers such as Richard Rorty [274] or Chantal Mouffe [232], disagreement is part and parcel of a democratic society where we must navigate these tensions in hopes of finding places of commonality from which to move forward.

We are now in an age of *machine action* where algorithms can benefit some individuals but may do so at the cost of harming others. Thus, we must not take the responsibility of implementing fair ML systems lightly. Transitioning fragile and contentious matters of human judgment to trained models must be done with care and forethought. As things stand, any engineer with a data set may codify a notion of fairness into an ML system without allowing for any disagreement or community consensus. In order to avoid slipping back into an essentialist morality where a small elite group decides what makes a machine action fair with no recourse, it is critical we expand the available

framings and considerations of "fairness" in ML.

Taking the stance that we should apply our modern ideas of fairness to ML, we offer the following proposition:

Proposition (P1): A machine learning system can only be fair with a contextual justification for the choice of a fairness construct and offering a channel for affected parties to actively assent or dissent to the fairness of the system.

In P1, a "fairness construct" is a definition for what fair is taken to mean in the problem space and the approach used to codify and measure that definition in training and application. What P1 implies is that a team implementing an ML system should have a sense of the viewpoints around fairness in a domain and be prepared to take a stance when choosing an approach. Further, it dictates some mode of disagreement be available. For example, either a) creating a method of algorithmic due process [96] where the fairness of a result can be scrutinized or b) working with an active community to address and resolve disputes over time.

A real-world example where we may require such a definition of fairness comes around building classifiers for predicting mental health issues using social media data. In research, there is a tacit acceptance that using social media data to predict mental health is fair [282, 216]. Models and outputs for this have been peer-reviewed and the system itself appears to pass as fair, or at least acceptable. Adopting the normative standard set by this work, one may believe it is fair for anyone to build a similar system to predict mental health, no matter what. And this is exactly what Facebook did–it was classifying teenagers by their psychological vulnerabilities such as feeling "insecure," "worthless," or "needing a confidence boost" [301]. However, Facebook's practice caused a lot of backlash from its users. The reaction to what such a system looks like in practice should raise a red flag that our questions about fairness must go beyond "was the technical approach fair" or "are the results are fair." It is our belief that the community of ML researchers and practitioners should also be asking questions such as "in what context?", "with what dataset?", and "with what objective?" should a system be trained to classify mental health. It is here we might consider that the discussion of Fair ML cannot be constrained to whether the classifier responds equally to

similar inputs. It is our position that ML practitioners must not skirt responsibility in questioning the ethics of what they build so long as some minimum equality criterion is met.

2.2.3.1 Contextualizing Fairness

In order to clarify our position on the role fairness should play in ML, we offer up three categorical questions one might use to frame an inquiry around the fairness of an arbitrary ML system: **(Q1)** "Is it fair to make X ML system?"; **(Q2)** "I want to make X ML system, is there a fair technical approach?"; and **(Q3)** "I made X ML system, are the results fair?". Each question requires a different set of considerations to arrive at an answer.

Q1 is asking whether a particular problem, in general, is fit to be approximated or decided upon by an ML system. Using P1, this kind of question requires us to consider the sentiments of the communities it would effect and whether we have a sense of what a fair automation of the task would look like. An immediate response may be, "well anything is worth a try if there's a reasonable data set and approach," but we will come back to why that is a specious assumption.

Q2 relies on a series of methodological questions perhaps best suited for experts. The fairness of an approach might rely on knowing what a good sample space might look like and the potential biases in historical samples. Answering this question would require an inquiry around whether the set of available features is a close and fair approximation of what we want to predict. Finally, we would hope a practitioner can weigh in on different trade-offs needing consideration for the choice of a training regime (e.g., [366, 191]).

Q3 gets at the question of treatment. That is, do the outputs of the algorithm actually correspond to what would constitute a fair response by a human. Answering Q3 requires us to run tests against the model and unpack the algorithm to determine qualities such as what input features were considered important, whether or not different groups have equal chance for misclassification, and whether any variables were acting as proxies for protected variables.

In light of these multiple framings, we take the position that no single framing nor problem space should dominate the realm of what we may consider fair ML practice. While we recognize

there will always be organizational ethics that must be further considered beyond what an ML practitioner can influence, we side with an interpretation of engineering ethics similar to that espoused by Langdon Winner [344] that the technology artifact itself has politics and is thus appropriate for normative evaluation. That is, he rejects that all ethics around technological artifacts are socially determined and argues that history has shown us that the artifacts themselves carry political and ethical weight. We now offer a few reasons why the advent of machine learning may embolden such a stance.

2.2.3.2 Why "Fair" Matters

In line with Winner's position, we should not see an ML model as a blank slate that can only be evaluated after appropriation by some organization. Not only might an ML artifact mediate ethically charged situations, but further it is an artifact carrying some amount of agency. From this vantage, "fair" ML is a recasting of the very idea of fair action in the human sense. Though it may be perceived that adding a "human in the loop" could solve our issue of ethical machine autonomy, research shows that "users may be prone to place an inordinate amount of trust in black box algorithms that are framed as intelligent" [303]. Meaning that even when an ML model is not acting autonomously, it is causing normative sway that is not neutral.

Further, we must consider that as actions coordinated by an ML model intervene in more of our lives, these actions are not always welcomed or requested [321, 25, 55]. As Frank Pasquale points out in "The Black Box Society," how we are categorized through data affects how we will be treated [254]. Grounding this thought, we ask whether someone who has never requested therapy, counseling, or any sort of diagnosis should be considered "open game" for a mental health classification. There is not an easy answer to this which begs the original question of whether any ML model that classifies mental health on social media data is fair.

Finally, we want to point to the fact that often the choice of a training objective is contentious in and of itself. There are some cases where the objective has a clear consensus, say in the case of classifying radiology images by whether a cancerous tumor is observed. There is little dispute that

the goal here is to be as accurate as possible at identifying cancer. On the other hand, determining whether an individual convicted of a crime might be a repeat offender is likely to solicit a lot of disagreement. Re-appropriating terminology from Richard Rorty [275], we distinguish between these two cases as *normal* and *abnormal* objectives, respectively. Meaning the objective of some ML tasks have a very clear grounds for consensus (normal) and others are highly disputable (abnormal). Due to the fact that ML could be applied to either kind of task, we believe this consideration elevates the level to which an ML model may be considered political or up for normative evaluation.

2.2.4 Current State of Fair ML

In light of our argument that questions of fairness operate contextually and that the advent of ML elevates our need for normative evaluation of technology artifacts, we move on to apply our contextual framings of fair to the contemporary Fair ML literature. We believe progress has been made, but mostly within the scope of framings Q2 and Q3 and almost exclusively employing some variant of "preventing disparate impact" as the definition of fairness.

2.2.4.1 Is it Fair to Make X ML system?

Outside of broad critical reflections on the use of technical systems [70, 142], there is yet to be much work characterizing grounds for why an ML system may or may not be a fair approach for a particular problem space. One might anticipate, given a normative definition of fairness, that questions of whether any data could ever approximate certain problems would provide grounds for healthy cross-disciplinary debate. However, "If we can, we ought to," is often treated as an unstated premise for technological development. An exceptional case, is the contemporary debate around autonomous weapons where thousands of scientists have supported a total suppression of development in this area [145].

In the Fair ML literature, a recent example of researchers asking more general questions about fairness is [57]'s evaluation of the ethical implications of ML for autonomous experimentation. Recognizing that practitioners have largely ignored established human-subject research

guidelines laid out by the Common Rule and Belmont Principles, the authors argue that automated experimentation may cause harms from privacy-violating inferences and exposing users to less-than-ideal outcomes by being part of the experiment. While falling short of actually questioning whether automated experimentation should be allowed at all, they suggest the adoption of an external review process in light of the fact that it's intractable to obtain consent from each user due to the complexity of the systems being tested. One might construe their high-level question as "Is it fair to make automated experimenting ML systems?" and their answer as "Maybe, but we should have external oversight."

Though we would be interested in a broader debate about the fairness of research ethics using ML systems, the conclusion drawn supports a more normative evaluation of fairness in this realm. Specifically, the idea that external review may be needed hinges on ideas of authority and context; namely, should we give engineers blanket authority to experiment on users? Akin to P1 above, research ethics are such that a review board should check your experimental standards and participants must be given certain rights. Thus, we agree with the suggestion of external review and further that certain autonomous experiments likely should not be done. Through our lens we would argue many other areas (e.g., ML for biometric inference, ML for emotional persuasion) are ripe for debate around the limitations of what systems are fair to implement.

2.2.4.2 I want to make X ML system, is there a fair technical approach?

Surveying the ML literature, we see three categorical trends signifying answers this question. The first grouping is research related to interpretability or "white box models" [354, 326, 336]. That is, this body of research approaches fairness by developing training methods that aim to produce interpretable ML models. Whether the model is fair becomes a matter of whether an explanation of the results is interpreted as fair. This intersects with the EU's upcoming adoption of "the right to an explanation" law and researchers' calls for algorithmic due diligence [96]. In our view, this approach has the most affinity with our P1 fairness definition given that it opens up the possibility for models to be interpreted by various subjects (allowing for contextual considerations) and sets

forth future possibilities for recourse and disagreement. The limitations of this approach are 1) this approach is not yet feasible for more complex algorithms (ie, most current trends in the ML field) and 2) fairness is bound by the ability for a subject to meaningfully understand and act on the interpretation as the interpretability guarantees nothing about the fairness of the algorithm itself.

Our second categorization includes attempts to resolve disparate treatment concerns by developing statistical independence between predictions and protected categories. These approaches include methods to satisfy fairness by enforcing robust sampling across groups [84], minimizing the difference in misclassification rates across groups [359], and training models where protected variables are neither explicitly nor latently used [214]. All of these methods define fairness in relation to treatment across known protected classes and give authority to an engineer (or perhaps partnered legal advisor) to *a priori* determine which variables are protected. This is a limitation due to the fact that some problem spaces may not have obvious or measurable categories that deserve protection (e.g., should we target someone for prescription drugs at a inferred moment of vulnerability?). A further limitation to this class of approaches shown by ML researchers is that there are inherent trade-offs between a) well-calibrated models, b) parity between groups in the positive class, and c) parity between groups in the negative class [191].

A third category of work involves decision-making algorithms that attempt to construct fair metrics for optimal choices. These attempts start with the adoption of a quality function that evaluates decisions and then argue for differing approaches toward optimizing choices such as always making choices that minimize regret (ie, integral over time of difference between choice and optimal choice) [178] or enforcing that a choice never be made when a higher-quality one was available [173]. A major limitation of this work is that its baseline standard of fairness is baked into the choice of a quality function, giving a lot of subjective authority to an engineer, and leaving only long-term and short-term optimizations to consider, which can be quite hard to reason about.

2.2.4.3 I made X ML system, are the results fair?

This realm of work might be construed as "algorithmic damage control" given that the attempt is not to make a fair algorithm, but rather to develop post-hoc analysis methods that help discover what may be unfair about a black box model. One class of approaches, again premising fairness on prevention of disparate treatment, involves developing mathematical tools to determine whether features related to protected categories (either directly or through co-variates) are influential on the model [17]. A number of these have come in direct relation to the buzz around the problematic recidivism instrument [88, 355]. Given our P1 criterion, these methods, at best, may help an end-user assess whether disparate impact occurred given a set of known protected categories. However, at worst, if we do not know which categories to assess for disparate impact, they may allow a model to pass as fair while unfair biases are still present. In summary, it is a progress that we have ways to verify if an approaches satisfies a disparate impact constraint; however, we are still levied with the challenge of deciding what the constraints should be.

The other major theme in this area of fairness brings back interpretability, but instead of model interpretability, the aim is interpreting why a particular result was obtained [272, 132]. While this has been successful in certain cases where researchers had an intuition about what kind of internal representation the model may have been using, these interpretations rely on naive, simplified approximations of the model. That is, they are unable to interpret the model globally and instead regress on a set of features in a localized subspace. Again, interpretability satisfies certain normative criteria of fairness by giving some power to an end-user to understand a result. However, so long as we cannot interpret the results globally, it's unclear how much power of recourse one may gain from such an interpretation. Further, this approach only allows us to interpret results using a predefined axis, meaning if we must know what we are looking for before this method becomes useful.

2.2.5 Concluding Remarks

In the above, we introduced a normative definition of fairness for ML and evaluated it against the current literature. We showed that there are multiple possible framings of fairness that raise different questions about "what is fair in ML" and require different evaluative constructs. Our first summary remark is to point out the limited nature of focusing ML fairness on disparate impact. A corollary of prior research shows that disparate impact and accurate model-making are in fundamental tension [191, 88] and ideal fairness defined this way may even be impossible [141]. Models are discriminators and as such, adding constraints affects one's ability to make a good classifier. Further, we may consider why it is engineers have pressed forward with different disparate impact constructs without yet inviting dialogue with the vulnerable communities for which they are trying to protect.

This brings us to address the most critical takeaway if one accepts a normative criterion of fairness (such as P1): engineers must invite in vulnerable communities and independent advocacy groups to engage in dialogue around fairness. We are sitting on the cusp of a societal transformation where many human intelligent tasks will be transferred to machine intelligence. Exciting as this might be, engineers should be cautious to move too quickly and leave behind the populations who are outside of the networks of the academic and industry elite. If we want to preserve our democratic essence, it is mandatory we develop the standards of the machine through inter-community dialogue. While this conversation is at its beginnings, there are already groups such as the Council for Big Data, Ethics, and Society forming to address these needs. Expanding the number of organizations discussing ethics and working with outside communities, making "usable fairness" a requirement in the development of ML tools, and ensuring that universities and business are teaching fairness and ethics to young engineers will all be critical to legitimizing the authority of the "fairness" embedded in our ML systems.

Chapter 3

Framing Policy

3.1 Coding For Respect

3.1.1 Prologue

This is a foundational piece with respect to the broader dissertation. This section lays out a framework for thinking about ethics in light of new capabilities of machines using Big Data and strives toward a lens through which policy could be made. Working from a fundamental ethical position that values individual autonomy, the possibilities of machines that can "disrespect" humans is explored. Developing this position on autonomy is crucial for understanding work throughout the dissertation as it is an ethical underpinning to many of the other issues explored, providing a moral resolve to the ethical thinking that follows.

Further, in order to develop the argument, it becomes critical to explore circumstances between human and machines that are not yet understood. In order to articulate these ethical conundrums that straddle the line of reality and fiction due to their possibility in the future, narratives are used as placeholder cases for analysis. Thus, the analysis of the framework goes through a set of ethical case studies that are well understood in the literature and through real events, but then moves to a set of case studies that are fictional and speculative and pre-emptive toward the future. In my original framework, these were developed as **boundaries**, or narratives that explore the ethical lines that could be crossed by machines. Using these narrative boundaries, I am able to argue for ethical positions within my framework that are intelligible and cohesive with the

broader problem space. Importantly, it allows this ethical analysis to extend into the future rather than pausing and waiting for the future where the lines drawn in the framework have already been crossed by machine actors.

3.1.2 Introduction

In a world where machines are increasingly being used to act autonomously in the place of humans, we all have an interest in preserving an ethical character to autonomous machine action (hereafter "machine action"). Indeed, this is often part of the very reason why we want machines to act for us in the first place: to remove the human flaws, biases, and oversights which lead such actions to be unethical or otherwise harmful when performed normally by humans. However, focusing solely on the ethically negative action characteristics (such as bias) that machines **might** [141] be able to avoid overlooks the ethically positive action characteristics that machines must still act in accordance with. Otherwise, in replacing human action with machine action, we will simply be trading one kind of unethical action for another – and machine actions might end up being unethical in even more problematic ways than the human actions they are replacing.

Some work in this direction is already happening. The current literature of fair machine learning works to overcome the problem of machines inheriting bias from their training data [42]. Aiming to discover [17] and constrain [197, 359] algorithmic bias, however, limits our focus to solving a single ethical issue of preventing illegal discrimination in machine action. And while we know we like models that are interpretable [326, 272], it is still unclear what boundaries we should be putting on the behavior of intelligent machines. As we move into the future, it is going to be critical, both socially and legally, that we build intuitions about what constitutes an ethical machine action, and formulate guidelines that engineers, technologists, policymakers, and lawyers can keep in mind when designing, evaluating, and regulating machine actions.

Our recommendation in this paper is that we should look to **philosophy** to help build these intuitions and formulate these guidelines. Philosophy, moral philosophy in particular, has the resources we need to expand and hone our ethical vocabulary, and to develop interpretive metrics

for regulating algorithmic systems.

In this vein, our specific contribution here is to bring attention to the importance of the concept of **respect** to any adequate account of ethical action – that is, respect **for persons**, or treating others never simply as means but always as **ends in themselves**. Respect is widely recognized in contemporary moral philosophy as a crucial component of ethical action, and, as we argue, it should be seen as a crucial component of ethical machine action as well. As we demonstrate, respect stands behind many of the ethical aspects of machine action that we already recognize and care about, such as privacy, fairness, accountability, and transparency. However, work still needs to be done to know how to systematically apply the concept of respect to the case of machine action. In addition, machine action includes some further, hitherto unrecognized dimensions along which respect also needs to be preserved and promoted.

The structure of our paper is as follows: In §3.1.3, we present a brief philosophical introduction to the concept of respect, and explain why respect is so important to the ethics of human action. In §3.1.4, in order to know better how to apply the concept of respect to the case of machine action, we present an analysis of machine actions into their six most ethically salient components. In §3.1.5, pairing together the insights from the preceding two sections, we show how the concept of respect can be applied systematically to the case of machine action, and detail the specific ethical considerations that emerge from this application. In §3.1.6, we briefly indicate the upshots of this framework for technologists, policymakers, and other practitioners.

Through this discussion, we have the following broad goals: First, to provide a philosophical framework for thinking through and evaluating the ethics of machine action. Second, to clarify the some of the specific ethical questions we should be asking about machine actions. Third, to highlight some ethical dimensions of machine action that are not being discussed much at present, but which must be in order to ensure a fully ethical character to machine action.

3.1.3 The concept of respect

Our approach to the ethics of machine action takes as its starting-point the concept of **respect**. We choose to focus on respect for three reasons. First, respect is generally thought of by contemporary philosophers to be one of, if not **the**, fundamental concept of morality, from which all our other, more specific moral obligations follow. Thus respect cannot be ignored in **any** discussion of ethical action. Second, as we detail below, respect has been fruitfully applied to a variety of other real-world ethical issues. Thus there is reason to look to respect for guidance in the case of the real-world issue of machine action as well. Third, as we show in §3.1.5, respect captures many of the more ethical moral norms of machine action that we already know and care about. Thus a better understanding of respect will clarify the importance of these norms, as well as possibly recommend further, hitherto unrecognized norms for our consideration.

What, then, **is** respect? Though there are many different kinds of respect [100, 270], one kind in particular has been of primary concern to moral philosophers: namely, respect **for persons**, that is, the respect that each person is due simply in virtue of being a person, the respect that is due to all persons **as such**. The basic idea is that, although in certain contexts some people might deserve more respect than others (due to differences in position, status, talents, or the like), there is nonetheless a baseline kind of respect that **all** deserve, due to the distinctive moral status as persons that we all share. "Persons, it is said, have a fundamental moral right to respect simply because they are persons" [111, §2].

This concept of respect should not seem foreign; indeed, it is natural for us to think that we have special categorical obligations to all other persons [76]. Yet respect would not have the place it has today were it not for the work of eighteenth-century German philosopher Immanuel Kant. For Kant, all our specific moral obligations and duties flowed from one ultimate moral principle, the "Categorical Imperative", which in its second formulation (the "Formula of Humanity") he expressed as the command that one act so as to treat persons "never simply as a means but always at the same time as an end" [180, 4:429]. Whatever else this dictum means, it is a command to

respect persons **qua** persons [100, p. 36], and this basic idea continues to animate moral philosophers today [193, 346, 99]. Some contemporary philosophers, following Kant, have taken respect to be the basis of all morality [117, 113]; and while others deny this [139], all agree that respect is an important ethical concept. For the purposes of this paper, we need not take a stand on this issue; it is sufficient merely that respect is accepted as an important ethical concept, in light of which moral principles and judgments can and should be formulated.

Why, then, is respect important? Take the following simple but powerful example (adapted from [138, 318]). Imagine a doctor who kills one of her patients, who is perfectly healthy, in order to harvest that patient's organs and give those organs to five of her other patients, who are each about to die from a different kind of organ failure. In one respect, the doctor has promoted "the greatest good for the greatest number" (the well-worn slogan of utilitarian moral theories), saving five people while losing only one, whereas were she to do nothing five would have died and only one would have lived. Nonetheless, the doctor's action clearly strikes us as morally wrong, as the doctor has treated her healthy patient merely as a means to her other patients' health. If the doctor had respected her healthy patient and recognized his fundamental moral rights (in this instance, his right to life and to his own body), she would not have engaged in such morally wrong behavior.

So what, more specifically, does respecting persons actually involve? What does it mean, in Kant's words, to treat others never simply as a means but always as an end? Primarily, respect involves **refraining** from certain actions and behaviors, such as the exploitation, manipulation, and debasement of others, violations of their rights, and interference with their decision-making or self-governance. In other words, respect sets a "boundary condition" for moral action, demarcating a subset of actions that are absolutely morally wrong. Respect does not always dictate exactly what we **should** do, but it does always indicate what we should **not** do. This is why Kant refers to the Formula of Humanity as "the supreme limiting condition of the freedom of action of every human being" [180, 4:431].

In addition to this behavioral dimension, respecting persons also involves a **deliberative** dimension. Respect requires not only that one act so as to respect persons as such, but also that

one give consideration to persons as such in one's deliberations about how to act. Kant, and many other moral philosophers since, have emphasized the deliberative dimension of respect over the behavioral. One reason for this is because respect itself is generally thought to be, fundamentally, an attitude or state of mind, and thus is more properly revealed in deliberation than behavior. But a deeper reason is because the deliberative dimension is thought to explain the behavioral: when one gives due consideration to persons as such in one's deliberations, one will, as a result, generally act respectfully towards them [56, p. 210]. Thus the best way to ensure respectful behavior is by encouraging respectful deliberation. This will be an important point to keep in mind as we move forward.

Lastly, it should be noted that respect for persons is not just some general principle of abstract moral theory; it also has a number of real-world application contexts. In the past half-century, moral and political philosophers have drawn on the concept of respect to explain and justify the nature and importance of moral rights [131], equality [343, 51, 140], social justice [247], privacy [50], political liberalism [56], and multiculturalism and the politics of recognition [315]. Of particular note is how attention to respect transformed the field of biomedical ethics, as well as actual health care practice, by emphasizing the importance of patient autonomy, which is now widely recognized as a basic principle of bioethics, and which has provided an essential counterpoint to the traditional norm of physician paternalism [46].

3.1.4 An analysis of machine action

In the previous section, we introduced the concept of respect and explained its importance to the ethics of (human) action. Thus, if we wish to preserve an ethical character to machine action, we must ensure that machine action lives up to this normative ideal, too. Yet it is not immediately obvious how the concept of respect should be applied to the case of machine action, as machine action is in many ways different from normal human action. However, this dissimilarity should not discourage us. As we will see, with a rigorous enough understanding of machine action, applying the concept of respect to machine action becomes relatively straightforward. Furthermore, clarifying

the differences between machine action and human action will reveal dimensions of respect that are unique to the case of machine action and thus all the more important to explicitly attend to, as we will argue in §3.1.5.

To know better how to apply the concept of respect to the case of machine action, we first need a more fine-grained analysis of machine action. Here we present an original analysis, which we believe captures the most ethically salient components of machine action. However, we should note at the outset that this analysis is not intended as definitive or unassailable; others can and should modify or expand this analysis as they see fit. The present analysis is intended, rather, as a template or framework for the sort of analysis of machine action that we believe is needed to think more rigorously and systematically about the ethics of machine action.

We analyze machine action into six components, divided into two groups (Table 3.1). The first group encompasses what we call the "kinds" of machine action, or the types of action or sub-action that a machine intelligent system might perform. Here we identify three basic kinds: observations, classifications, and interventions. The second group encompasses what we call the "strata" of machine action, or the levels at which a machine action might be carried out. Here we identify, similarly, three basic strata: individual, collective, and iterative.

Group	**Component**		
Kinds	Observation	Classification	Intervention
Strata	Individual	Collective	Iterative

Table 3.1: The components of machine action

Some basic examples will help clarify what we mean by each of these different components. Let us begin with the three kinds of machine action.

"Observations" are actions that correspond to some mode of data capture [19]. These actions are often defined specifically by an engineer who designs a data mining or behavioral tracking system; however, autonomous experimentation [57] is beginning to replace the need for human definitions. Examples of observations are taking behavioral analytics, server logging, storing revision

histories, and sensor time series. Observations are, of course, a component of human action as well; but notably, machine observation elevates the ability for tracking and experimenting beyond what a human could do. Furthermore, whether constructed by a human or an autonomous algorithm, machine observation establishes the space of possibilities, or ground truth, that will determine all further actions that the machine or any associated machine intelligent systems will take.

"Classifications" are actions that assign users classifiers and change how they are "seen" or treated by a machine intelligent system (typically, on the basis of data that has been collected from a prior observation action). Classifications are the process of using statistical and computational methods to cluster, organize, or label users and their data [304]. Examples of classification include training a classifier that predicts a user's religion [194] or mental health [269, 216], the results of which will then change what ads or content the user sees. Trained models may be construed as a species of classifications, as such models form the basis from which future decisions are made. The act of creating classifications allows for the further embedding of those classifiers or models into any relevant autonomous machine system.

"Interventions" happen when a machine actually does something, changes something, or interacts with a user directly (typically, on the basis of previous observations and/or classifications). Interventions are the positive product of a machine acting in the world. Examples of interventions are establishing the visual order of an interface or suggesting auto-complete text. Interventions are the most material and manifest action that a machine may take. The results of an intervention are felt by, and may even harm [321], the user, whether or not they comprehend its occurrence or consequence.

By combining these three kinds of machine action together, we can form a pipeline where each kind of action feeds into the next. Thus an observation can create data, which can then be categorized or structured via some analysis or training, which can result in a model or set of categories which allows for a specific intervention on a user. A simple example of this would be the filtering of social media feeds: In this machine intelligent system, data is first captured on the basis of user clicks or likes of content (the observation actions); next, data that has been collected

from many users goes into the training of a system that ranks content given a specific history (the classification action); and finally, an autonomous system filters and orders live content, using those learned ranks, at the moment a page request is made (the intervention action).

Here it should be noted that, although interventions may seem like the purest and most important form of machine action (since they are the finite moments when a human consequence is actually felt), each kind of action in this group uniquely impacts the conception of and behavior towards the user in a machine intelligent system. This point is especially relevant when applying the concept of respect to machine intelligent systems, since, as was noted above, respect involves not only a behavioral but also a deliberative dimension. In this regard, the steps that feed into and inform any machine intervention (i.e., the machine's "deliberations") are just as important to respect as the intervention itself (i.e., the machine's "behavior"). We return to this point below, in §3.1.5.

Let us now turn to our second group of components, the three strata of machine action. Beyond the **kinds** of actions that machines may take, there are also separable **loci** at which machine actions take place. The "individual" stratum considers a machine action from the perspective of its impact on a single user. Examples of individual-stratum actions include observing a single user's history of responses to a marketing campaign, assigning a single user a category relevant for targeting, and rendering a specific article or ad at the top of a user's feed.

The "collective" stratum views machine actions from the perspective of how they affect a population of users. Examples of collective-stratum actions include collecting data from all employees in a workplace; a data classification that (directly or indirectly) serves as a proxy for a protected class, such as race or political orientation; or an image captioning system that exclusively mistags black faces [5]. To be sure, collective actions emerge out of the combination of many individual actions; nonetheless, they warrant separate ethical consideration, as the combination of individual actions is itself a matter of ethical concern. This is especially true when the combination of individual actions results in impacts linked to a shared attribute that targets a specific community (e.g., disparate impact [42]).

The "iterative" stratum is the most distinctive component in our analysis of machine action. Iterative actions are actions viewed from the perspective of the dynamics and impacts that occur due to same basic action being repeated and reiterated numerous times (as, indeed, many machine actions are). We separately identify this stratum due to the fact some actions, though harmless when looked at in isolation, become disconcerting once the action is performed repeatedly and continuously. Take, for example, the persistent shaping of a user's news feed. If we were to analyze the personalization of a news feed on the individual stratum, we may find nothing wrong with it. But when we consider the same personalization being iterated over time, to the extent that a user sees nothing but what is personalized for them, the psychological and social consequences of the action become tangible.

The preceding discussion should suffice to clarify our six-component analysis of machine action. If it is not already clear, any machine action can be described according to both its kind and stratum. Thus, by treating our two groups of components as axes and crossing them, we can create a 3x3 matrix, in which we can locate any particular machine action according to its kind and its stratum (Table 3.2).

This analysis of machine action may seem illuminating in its own right, but recall that our ultimate reason for introducing it was so as to know better how to apply the concept of respect to the case of machine action. And indeed, as we will argue presently, for each kind and stratum of machine action that we have identified, there is a distinct and specific moral obligation, arising out of the fundamental moral obligation to respect persons, which must not be violated if the action is to be ethical. (Note that this is a necessary, but not necessarily sufficient, condition for ethical machine action.) Understanding these different obligations is, we believe, foundational to the structuring of any future regulation that looks to protect humans in the face of actions taken by autonomous machine systems. In the next section, we outline what each of these obligations are, how they derive from the more basic obligation to respect persons, and explain why they are important and relevant to the case of machine action.

3.1.5 Respect in machine action

Recall that, when applying the concept of respect to the evaluation of any action, the basic question we must ask is: Does the action respect the persons involved in and/or affected by the action? (Does the action treat them simply as means, or as ends in themselves?) Accordingly, when applying the concept of respect to the evaluation of machine actions, the basic question we must ask is: Does the machine action respect the users involved in and/or affected by the action? (Does the machine action treat the users purely as instrumental to some goal, or as autonomous individuals with the basic moral standing of persons?)

As stated, this basic question is still too general and abstract to offer any specific guidance in the evaluation of machine action. However, more specific and focused questions can be posed, in light of the analysis of machine action introduced in §3.1.4. This is because, as we argue in this section, respect is manifested in a distinctive way for each kind and stratum of machine action defined above. These results are summarized in Tables 3.3 and 3.4 below. But to more effectively introduce and convey these ideas, we first present a number of concrete cases that illustrate the different forms that respect takes for each of our kinds and strata of machine action. These cases will also clarify how prior work in machine learning and computing ethics fits into our framework.

A preliminary note: One of the distinctive aspects of our framework is our assumption that respect must be preserved and promoted for **all** kinds and strata of machine action. This assumption may strike some as surprising, as it may seem that, when it comes to the ethics of machine action, it is only machine **interventions** that we should be worrying about, since this is where the discrete harms of machine actions actually materialize. However, here again we point to the fact that respect for persons involves more than the here performance of respectful **behavior**. In addition, it requires manifesting respect in one's **deliberations**, or the considerations and sub-actions that lead up to and result in one's behavior. Thus, it is crucial to recognize that we are here outlining a matrix for discovering the **provenance** of a machine's disrespect for its users, as opposed to identifying specific machine interactions that actually cause harm. Comprehending this broader

spectrum of ethical action is, we believe, the only way to tease apart the increasingly complex interactions that machines are having, and will continue to have, with humans.

That being said, let us turn now to our cases. We present one case for each of the nine regions in our matrix (Table 3.2). We begin, in §3.1.5.1, with some familiar cases of machine action (Cases A–D), and then turn, in §3.1.5.3, to some emerging and less familiar cases (Cases E–I).

Strata \ **Kinds**	Observation	Classification	Intervention
Individual	A	C	F
Collective	B	D	H
Iterative	E	I	G

Table 3.2: Cases in our matrix of machine action

3.1.5.1 Familiar cases

Some of the kinds and strata of machine action in our analysis already have a rich literature concerning the ethical problems they raise. Here we briefly detail four of the nine regions in our matrix, whose ethical issues are widely recognized in machine learning and computing ethics. Yet as we demonstrate, these well-known ethical concerns can all be seen to derive from the basic moral obligation to respect the persons (or users) involved. This, in turn, provides some initial justification for our respect-based framework.

Case A: Private data collection

[Individual Observation]

Example: Devices and applications that process personal information about users often violate reasonable expectations of privacy. The use of technology to collect or show private information at unexpected, unknown, or unwarranted junctures has already occurred in many cases. The growth of the internet of things, and ubiquitous computing more generally, gives rise to machines able

to gather information beyond what an individual can comprehend or manage. This threatens our individual ability to manage what is known about us and to whom. A 2015 study of mobile applications [360] showed that personal information such as location and email address are being shared with third-parties with no notification to the user. The FTC recently settled a case with Vizio for not disclosing how it was collecting user information [13]. Uber's God View [54] further proved that without reasonable checks, the data we expect to be private could be reappropriated outside of our control.

Analysis: There are two clear ethical concerns in this case. First, such data collection violates the user's rights, namely, their right to privacy. Second, such data collection is (often) opaque to the user. Our framework explains why these are legitimate ethical concerns, for both can be seen as failures to respect the user. Moreover, both failures can be seen to correlate to the kind and stratum of such actions, that is, the fact that they are individual observations. First, for any individual action to respect the user it affects, it must refrain from violating that user's fundamental human or moral rights, that is, the rights the user has simply in virtue of being a person [131], and these include the right to privacy [50]. Second, for any observation action to respect the user it observes, it must ensure that the observation is transparent, such that the user can be aware of it. Observations that fail in this regard end up treating their users purely instrumentally, as mere data points to be collected, and not as the persons that they are.

Case B: Uneven data collection

[Collective Observation]

Example: Deploying a data capture system that does not collect data evenly across populations creates the problem of over- or under-sampling certain populations. We expect the data we collect to represent the true state of affairs, eliminating the biases that humans carry. However, we are finding that technology-enabled data capture carries its own potential to exacerbate inequality between different communities [39]. The escalation of predictive policing practices has raised serious

questions about the fairness of the impacts. Attention has already been paid to the fact that facial recognition databases disproportionately represent black faces [300]. Conversely, data collected to solve problems might bias solutions to aid only those with the most access. Such was the case with the StreetBump app which helped get potholes get fixed, but due to faulty data collection methods, did so primarily for richer neighborhoods [95].

Analysis: This case introduces a further ethical concern, namely, that such data collection seems unfair, in that it disproportionately and unequally affects the users in its population. Again, our framework explains why this is a legitimate ethical concern, as this, too, can be seen as a failure to respect the users involved. Furthermore, this concern arises from the fact that, in this case, we are considering data collection at the collective stratum, that is, from the perspective of how such actions affect a population of users. For any collective action to respect its users, it must ensure that they are all treated fairly and given equal consideration [343, 51]. This is because all users share the same fundamental moral status as persons. Flagrantly unequal and unfair treatment of user populations fails to respect this shared moral status.

Case C: Triangulating sensitive information

[Individual Classification]

Example: When personal data is handed over to companies, users can be unaware of what they are truly disclosing. The ability to predict undisclosed attributes about individuals using data they provided has already been proven [194]. There have even been cases of companies selling highly-sensitive, anonymized medical data to companies that can de-anonymize the data set using complimentary data [317].

Analysis: The ethical concerns in this case are slightly more subtle. Like Case A, part of the issue is that such actions seems to violate the user's right to privacy [96]. This is not surprising given our framework, since this case, like Case A, is considering action at the individual stratum; and as we saw above, respect at the individual stratum requires not violating the user's rights. In addition,

the actions of this case seem objectionable because the user has not been given a voice, or say, in their classification; rather, the classification was made on their behalf, without their consultation or consent, and without even the possibility for the user to object or intervene. Again, our framework explains why this is a legitimate ethical concern, as such "voiceless" classification of a user can be seen as a failure of respect. Indeed, for any classification action to respect the user it is classifying, it is essential that the user be given some degree of voice in their classification. Otherwise, the classification ends up treating the user paternalistically, as a mere means to some desired data-set, and not a free and self-determining individual.

Case D: Sentencing algorithms

[**Collective Classification**]

Example: The highly-publicized case of the COMPAS recidivism scoring algorithm highlighted the threat of machines classifying people along racial lines [88, 27]. Once machines codify relationships in high-dimensional feature spaces required for complex models, they have the potential of constructing unfair classifiers that discriminate on the basis of identity. Further, the use of a machine system to solve a contentious human problem minimizes the affected person's ability to understand and potentially redress any harm.

Analysis: This case raises two ethical concerns, which we have already seen above. As a collective action, we must ensure that the action is fair; and as a classification action, we must ensure that the action gives users a voice. On both counts, this case seems problematic, treating the users in its population unequally, and not giving its users any say in their classification.

3.1.5.2 Summary of familiar cases

These familiar cases highlight the relevance of respect to the ethics of machine action, and detail how respect can be more thickly conceptualized within our framework. To ensure respect, individual actions must not violate the user's **rights**, collective actions must treat all users **fairly**

and **equally**, observations must be **transparent**, and classifications must give users a **voice** (Table 3.3).

These specific norms and considerations should already seem familiar to those in the machine ethics community. Yet our discussion highlights that these acknowledged ethical concerns derive from the basic moral obligation to respect persons. In other words, in emphasizing privacy, fairness, accountability, and transparency, practitioners in the machine ethics community have already tacitly been working to ensure respect in machine action.

This is significant, because respect manifests itself in the other dimensions of machine action, as well, in ways that are **not** at present being discussed. A complete ethics of machine action must also pay attention to these forms of respect.

3.1.5.3 Emerging cases

Thus far, we have looked at cases of individual, collective, observation, and classification machine actions. What remains to be discussed are cases of iterative and intervention actions. Such cases are especially relevant within our framework, since respect in these cases is not as well understood as it is in the others. As we argue below, respect in these cases involves preserving and promoting the user's **autonomy**: the user's fundamental right to act freely and of her own accord, rather than being coerced or having decisions made for her.

To further comprehend the threat of these less-understood issues, we present novel case studies meant to aid in our collective understanding of the dimensions of respect these areas of our matrix put into question.

Case E: Low-level behavioral tracking

[**Iterative Observation**]

Example: Elaine uses her mobile phone to read news. Even though she chooses not to click on certain articles, she sometimes pauses, shocked by the headlines. To try to minimize the amount of shock content she receives, she often copies and pastes information to a separate application to

fact check. While she believes to be avoiding giving information about what shocks her, the fact observation continues across all facets of an interaction, she cannot help but inform an intelligent machine of the truth via her behavioral metrics. However, without some other feedback mechanism for Elaine, it is possible, if not likely, she will be fed more shocking content.

Analysis: This case is an example of an iterative observation: it is an observation, because data is being collected; and it is iterative, because the data collection is done expansively and continuously. Like other observation actions, part of the problem here is that the data collection is not transparent to Elaine. But this case also has a distinctive problem due to its iterative dimension. Despite Elaine's best efforts to prevent the app from knowing too much about her habits, the app's iterative data collection outstrips her ability to do so. In this way, the app undermines her ability to act freely and of her own volition, and in this regard compromises her autonomy.

Case F: Targeted advertising

[**Individual Intervention**]

Example: Frederick wants to go to Harvard for college. Leading up to receiving his admission decision, he displays anxious behaviors by scouring the internet for information about who has received admission and posting messages to his friends showcasing how much the uncertainty is bothering him. Adapting to his social media data and behavioral metrics, he begins receiving information related to anxiety and depression medication. Despite never having thought of himself as depressed or feeling unhealthy anxiety before, upon being rejected from Harvard, in a moment of vulnerability, he is shown an ad framed as "Is anxiety harming your performance?" and chooses to click and purchase.

Analysis: This case is an example of an individual intervention. Its distinctive ethical concern comes from its intervention dimension. The machine is interacting with Frederick in an attempt to influence him and make him respond. In itself there is nothing wrong with a machine (or a human, for that matter) doing so. However, because the machine is using its trove of data to show

Frederick the ad precisely when it knows he is at his most vulnerable, the intervention can easily seem coercive or exploitative. If a human were to intervene in this manner, we would consider it predatory. In its specificity, the intervention compromises Frederick's autonomy and disrespects his status as a free, autonomous individual.

Case G: Influential advertising

[Iterative Intervention]

Example: Just before bed, Gina tends to browse the internet for new clothing. Recently, she has noticed more photos and ads with fashion models showing up across all of her platforms, especially at night. Gina is now beginning to feel insecure when she's tired and can't stop being bombarded with expensive clothing and thin models. After a few nights where her mind began to wander toward very negative thoughts, she decided she should stop looking at clothes before bed and read something else instead. However, now that her interests and browsing times have been learned, even on unrelated platforms, she is continuing to be shown models in every ad bar. Her insecurity worsens and she is now considering counseling for body image issues which she never used to have.

Analysis: This case is an example of an iterative intervention, combining elements from Cases E and F. As we have already seen, both iterative and intervention actions run the risk of compromising a user's autonomy. In this case, the iterative dimension is what is particularly ethically problematic. As a one-off intervention, showing an ad to a user does not disrespect their autonomy; we can reasonably expect the user to display some resilience. Yet when performed iteratively, this intervention becomes more of a concern, as it can shape, influence, and determine a user's self-conception, and thereby compromises a user's ability to form that conception freely and for themselves. These tactics have already been discussed as an emerging possibility in targeted marketing [231].

Case H: Filter bubbles

[Collective Intervention]

Example: Heather and Henry are friends in real life and social media. Though they have very separate interests, their friendship is built on a mutual respect for one another. One day Heather sees information in her news feed about a new federal policy proposal that she finds concerning for her family. Knowing that her friends may care to know how this could impact her, she posts a heartfelt plea to stand against this policy, outlining the impact it would have on her particular family situation. She asks that if you disagree not to bring it up in front of her brothers or sisters who are having issues. Meanwhile Henry receives news that explains a polar opposite view of the same policy and his feed filters out Heather's plea due to distinct political interests. Doug then raises the issue in front of Heather and her sister while at a party, which hurts Heather given her public plea. Neither knows what the other has seen and thus both assume the other is completely unreasonable.

Analysis: This case is an example of a collective intervention. The intervention itself (like Cases F and G) is concerning due to the fact that it is opaquely filtering information in a way that restrains individuals' – such as Henry's and Heather's – ability to freely communicate. Moreover, this case is a collective intervention, due to the fact that individuals are being clustered along interest lines and the effect of the intervention is to segment the user population. Thus, while the personalization of information feeds is not always in itself problematic, in this case it both limits the communicative capacity of the individuals and compromises the assumed equality of public dialogue by restraining information flows across groupings. Critically, the fact Henry and Heather **want** to communicate with each other in a public setting, yet are being **opaquely** undermined in doing so, disrespects their autonomy.

Case I: Collateral classification

[Iterative Classification]

Example: A large online search engine begins classifying words and semantics that are likely to signal highly politicized or fake news. Leading up to a major federal election, Isaac is following a

surging fringe candidate building conversation around universal basic income. Though the topic is getting traction, it does not get taken up seriously by major media networks. Beyond covering this surging candidate, a number of smaller publications have begun publishing articles using unsound claims around environment, refugees, and medicine. Seeing results being overcrowded by fake news, the search company decides to apply its classifier to its search engine to improve people getting news with sound sources. With only a few days until the election, all articles around universal basic income along with those discussing unsound science disappear from the top of search results. By the time the company realizes the issue, the election day has come and went and the surging candidate lost traction and exposure.

Analysis: This case is an example of iterative classification. It is a classification because we are dealing with a trained model classifying news content that is likely to be fake or biased; and it is iterative because the same internal logic of the model is being applied over and over again to all articles. In this case, the model was very good at identifying fake news and lowering its rank in search results. However, it came with a hidden cost of misclassifying articles representing real, factual political discourse that carried features intertwined with those signaling fake news. Yet the ethical concern in this case goes beyond the model's failure to accurately classify real news. More fundamentally, the model impeded voters' ability to find news related to their issue of choice [335]. It decided what was relevant for them, rather than allowing users to freely make this decision for themselves.

3.1.5.4 Summary

In this section we have argued, on the basis of the above cases, that the general moral obligation to respect persons can be seen to take on a more specific form according to the kind and stratum of machine action at issue. Thus, each kind and stratum can be associated with a specific norm of respect (Table 3.3), and each of these norms can be understood in light of a corresponding ethical question (Table 3.4).

Group	Component	Norm
Kinds	Observation	Transparency
	Classification	Voice
	Intervention	Autonomy
Strata	Individual	Rights
	Collective	Fairness
	Iterative	Autonomy

Table 3.3: Norms of respect

Norm	Question
Transparency	Is the user (able to be) aware of the action?
Voice	Is the user able to influence the action?
Rights	Does the action violate the user's rights?
Fairness	Does the action treat all users fairly and equally?
Autonomy	Is the user still able to act freely and of her own accord?

Table 3.4: Questions of respect

Here, then, is our basic recommendation: If you are considering the ethics of any particular machine action, first identify which cell(s) in our matrix the action falls under, and second pose to yourself the corresponding questions of respect.

Admittedly, this recommendation leaves many of the hard questions unanswered. We do not assume that it is obvious when, for example, an observation is sufficiently transparent, or a user is able to act freely and of her own accord. Yet this is not the purpose of our framework. Its purpose, rather, is to clarify the kinds of questions that we should be asking when evaluating the ethics of machine action. By having these questions in view, it is our hope that the task of designing ethical machine actions will seem more tractable and a little less daunting. The hard work of answering these questions, however, lies with all of us.

3.1.6 Concluding Remarks

This paper has addressed the ethics of machine action using a philosophical framework based on the concept of respect. Our fundamental assumption has been that a linchpin of all ethical action is that persons – the users in intelligent machine systems – must be treated as persons, that is, as autonomous agents and ends in themselves. Based on the multifarious abilities of intelligent machines, we applied our framework of respect to clarify the various ways in which machine actions can undermine this fundamental moral obligation and disrespect their users, by treating them as purely instrumental to their objective. We believe this framework encompasses much of what has already been looked at through the lenses of privacy, fairness, accountability, and transparency, while at the same time expanding our intuitions about ethical machine action in emerging cases.

Our framework of respect is a pathway towards what could ultimately guide organizational policy, engineering best practices, and government regulation. When building systems that interact with humans and their data, a first evaluation should be to ask, "What is the conception of the user in this system?" Namely, "Does our system treat the persons involved as ends in themselves, or purely as means to the system's objectives?" If we cannot meaningfully answer these questions, then we may be designing a system that unethically instrumentalizes and dehumanizes its users.

Researchers are already undertaking the challenge of quantifying and systematizing definitions of "fairness", "accountability", and "transparency". We believe our rubric may aid in helping practitioners know when they should be applying these tools. For example, our framework clearly distinguishes between systems whose objectives are normatively contentious, insofar as their actions impinge on their users' rights and autonomy (e.g., passive data capture mechanisms) and systems whose objectives are normatively unproblematic, insofar as their actions do not violate their users' rights or autonomy (e.g., screening medical imagery for cancer).

We encourage researchers to further clarify and expand the kinds and strata of machine action we have proposed. Many of these cases, particularly the iterative stratum, involve complex social dynamics that will need more robust definitions as new cases emerge. There is also work to be done

in designing better protocols for transparency and systems analysis in terms of the interpersonal dynamics promoted for affiliated users. As we have seen with fairness, once we can build robust conceptions of what ethical treatment looks like, we can move on to quantifying and implementing approaches. One major recommendation of our framework is that future systems should also be designed to give users a voice, by, for example, receiving live feedback from users and incorporating this feedback into future actions.

Admittedly, this paper does not solve the ethical problems we've highlighted; but this was not our aim. Our goal was to clarify our intuitions around how machines may curtail the rights and freedoms that we all see as ethically fundamental.

We hope readers walk away with a better understanding of the kinds of questions they should be asking of intelligent machine systems and an improved comprehension of the ethical space that machine actions occupy in human life. Given our current discourses around rights and regulatory mechanisms to protect those rights, we believe a framework like ours can take us further towards a systematic approach to the ethics and regulation of machine action.

Chapter 4

Ethical Thinking within Computer Science

4.1 Ad Empathy: A Design Fiction

4.1.1 Prologue

The following section uses design fiction to create, what I call in my framework, a **target**. That is, this design fiction elaborates on a fictional product that can be conceived as possible, or even probably, given research and market trends. From the lens of this product, a world of ethical questions come alive. This particular fiction is exploring the line between manipulation and personalization via emotional targeting using AI. Instead of arguing for a particular boundary like other work, the piece forms a fixed position in the future where this product exists and we now have to deal with what that means. The piece also calls back to my earlier-developed notion of autonomy in the problems that it raises. The hope is that the piece triggers awareness and offers engineers a point for debate and ethical thinking about a product that is becoming eerily close to existence.

4.1.2 Product Introduction

Today's competitive attention economy requires brands to reach customers in personal and affective ways. Years of research and experience establish that personalization is effective for ad targeting and affecting user and consumer attitudes [20]. However, personalization is also now a saturated approach. The relative ease of obtaining consumer preference data makes it common

for online advertisers to know what a customer wants. Companies wanting the competitive edge now need to know when a product is best advertised and how it should be framed. Knowing this demand, we are happy to launch Ad Empathy, an AI marketing solution supporting brands to make emotion-sensitive marketing decisions.

Our API Resources are designed to help our clients generate content for ad impressions, catering to the dynamic needs of the diverse individuals in their audience. We work with most major social media platforms and search engines to create connected profiles of customers that can be accessed from any ad client via the Ad Empathy API. For each advertising platform you would like to integrate with Ad Empathy, simply add your company's registered OAuth tokens using the Ad Empathy Dashboard and within 48 hours we will have trained models for each of your customers and customer types. From that point onward, you can use the Ad Empathy API to design your ad impressions on any connected platform. To use Ad Empathy as a full-cycle marketing platform, you may also register your product inventory with our platform to track emotional responses to product-specific brand interactions and improve our models.

4.1.3 Getting Started

Before making any requests using our models, you should contact a member of our Sales Team to discuss pricing options or obtain a free trial. All API Resource requests must contain a valid token pair $< client - token >$ and $< client - secret >$, a $< cookie - id >$ for the user, and optionally a $< platform - id >$ to specify the ad client platform. Developers building platform-agnostic services can use our Accounts API to obtain valid $< cookie - id >$'s for building cross-platform ad campaigns and event triggers.

4.1.4 API Resources

Once you have obtained valid token pairs, integrated your external ad platform's tokens, and see the green check mark at the top corner of your Ad Sense Dashboard, you can begin using any Ad Empathy API Resource.

4.1.4.1 Mood

Mood

get - GET /mood/now/< cookie − id >

Returns current emotional state (mood) of user as a list of top ten moods by confidence

list - GET /mood/list/< cookie − id >

Returns a list of frequencies for all moods categories that Ad Empathy has related to the specified user.

Mood.product

list - GET /mood/product/< product − id >/< cookie − id >

Returns a list of product IDs and the mood that is most positively associated with a customer interaction.

Mood.topic

list - GET /mood/topic/< cookie − id >

Returns a list of content topics and our highest confidence mood association for that topic.

4.1.4.2 Trend

Trend

get - GET /trend/now/< cookie − id >

Returns the predicted emotional states, ordered by confidence, for upcoming 30-minute time interval.

list - GET /trend/daily/< cookie − id >

Returns a list of 30-minute time intervals over 24-hours with the most common emotional state associated to each interval.

4.1.4.3 Response

(API Resource available only to customers using Ad Empathy Trackers for their product inventory)

Response

get - GET /response/< product − id >/< cookie − id >

Returns the user's last cached online emotional response to an interaction with $< product - id >$.

4.1.4.4 Expression

Expression.text

get - GET /expression/single/< emotion >/< cookie − id >

Returns the syntax tokens most commonly associated with the user's online expression of the emotion.

list - GET /expression/all/< cookie − id >

Returns a paginated list of emotional states, sorted by their frequency, and the most common syntax tokens associated to that state.

4.1.5 How Does It Work?

Ad Empathy is a state-of-the-art multi-model AI ecosystem that leverages the volume and velocity of online behavioral data by training user-specific machine learning models. The core of the system is a Long Short Term Memory (LSTM) neural network trained specifically to predict the evolution of moods using temporally-structured data coming from online activities (eg., text from posts, click content, reactions to others' posts). Our company began training this model nearly five years ago when researchers first found Gated Recurrent Units as a solution to cutting through the noise of online data [15]. After years of fine-tuning and learning how to transfer models between different users and incorporate multi-modal data, we found we had sown the seeds of something much bigger than a mood prediction model. In short, this core model became the heart of a system of interacting models. Developing our expertise in model transfer allowed our team to take layers of our novel LSTM model and combine them with convolutional layers or other Recurrent, language-processing layers, and train them as Generative Adversarial Networks to blossom the wide functionality of novel content creation you see today.

When your company opens an account with Ad Empathy, our system begins by data mining all social media content and brand interactions available for your customer base. After mining all historical data about your customers, we place their user accounts into our reactive event loop that keeps tabs on new activities across any connected platform. Prior to training, we run all the data through a noise reduction network trained specifically to identify relevant emotional content. Using the filtered data set, we fork fresh versions of our base model and begin training a unique mood model for each of your customers. This training continues until the confidence of our predictions meets a certain threshold. Testing is done using a data set we capture and separate during the data-mining phase. Our central model (the one underneath the Mood API) takes in time-structured online activity for a user and outputs a likely current mood given the most recent observation. This model is then transferred into our second network, which chunks your users' history into 24-hour segments and trains a model that predicts the upcoming 24-hour emotional cycle (and provides the backbone of our Trends API!).

Once we have accurate models for our Moods and Trends API, we do fine-grain analysis on specific data such as text and photos. This process starts by performing a topic-modeling analysis on all user text and browsing history to break up each user's' history into topic-specific data sets. Further, each user photo is analyzed for facial expression, object detection, and captioning to develop visual insights into the personal aesthetics of your customer's emotions. A core value that Ad Empathy offers is recognizing that each product a customer purchases is embedded in a different context and thus requires a different cognitive model to understand underlying emotional relationships. We develop those models along many dimensions that account for complex relationships between emotions and brand sentiments.

Important to understanding how Ad Empathy works is that each API your team uses is operating with different custom models and parsing techniques that branch out from of our central mood-recognition network. Our Expression API, for instance, uses sentiment analysis in tandem with a generative adversarial network to parse user text and then learn how to generate novel text that expresses the same sentiments while staying within the known vernacular of your customer.

The adversarial network is trained against the core mood model, which allows rapid exploration of the syntax space observed and parsed from your customers' online platforms.

If your company would like to learn even more about the inner-workings of Ad Empathy, feel free to make an appointment with our Machine Intelligence Team to discuss specifics or let us know how you think we could improve our process.

4.1.6 Example Use

Working with customers, we have found solutions that mix and match our APIs to help you generate the relevant content and design marketing campaigns most appropriate to your products. We explain some of our most successful applications below:

4.1.6.1 Time Cycling

Our research has shown that many customers have predictable emotional response patterns based on time of day. It is often reliable that a customer will elicit more positive emotions to food around 11AM; however, this response will diminish leading up to around 2PM as it becomes more likely they already ate lunch. For this reason we recommend time cycling campaigns for products with emotions that are highly correlated to temporal patterns. For this, we recommend analysis of your products with our Trends Resource to discover your most temporally stable products and to make inferences about how they are associated across time. Then using our Expressions Resource, you can design context-sensitive Content Ads that can portray your product regularly at the times associated to the emotion best suited for your product.

4.1.6.2 A/B Emotional Testing

Not sure whether your product is better fit to when your customer feels happy or angry? Try A/B Testing emotions instead of features. Combining our Impressions and Response APIs, your team can try your ad impressions against different emotional conditions to see what elicits the most positive response. This can improve how you understand how your product is being perceived and

better inform our models.

For well-modeled user profiles, your team may try running simulations using our Impressions and Expressions APIs. You can pilot your A/B tests, discovering correlations between ad impression and emotional responses and designing ad impressions with the right emotional language.

4.1.7 Appropriate Use of Ad Empathy

The purpose of Ad Empathy is to support businesses in employing emotional insights as they create online advertisements. We love seeing our customers rapid prototyping new ad campaigns and trying out new combinations of our models to maximize the utility emotions and timing play in your ad impressions. Ad Empathy, however, is not meant to be used as a research platform nor should it be used to target specific customers and invade their privacy. We do not approve of customer-specific analysis that exposes potentially sensitive vulnerabilities related to private dimensions of a customer's mental state. Ad Empathy should also never be used in relation to medical data or to support mental health inference relative to emotional trends. Similarly, our insights should remain in the realm of marketing and should not be used in decision-making algorithms related to employment, education, housing, or health. Though we are proud of the accuracy of the accuracy of our system, it is not appropriate to use such predictions to make firm decisions that could negatively impact your customers. If your company is focused on biomedical or employment-related inference, please contact our Customer Relations Team to discuss fair uses of data and how to access our models for purposes outside of our available products. Projects that are funded by a government agency should speak to an Ad Empathy representative before using our products. If your use of Ad Empathy goes beyond marketing, we offer consulting services to help your company develop an ethical and accurate system that incorporates emotional insights.

Thank you again for using Ad Empathy!

4.1.8 Author's Statement

The goal of this design fiction is to structure discussion around a technology that is at the cusp of creation, regardless of whether it emerges in this exact form. Industry demand for novel forms of personalization and audience targeting paired with research trends in affective computing and emotion detection puts us on a clear path toward emotion-sensitive technologies. With both the capability and economic incentives in place, we must, as a community, carefully define lines between what we consider fair marketing applications of technology versus unwelcome and unfair intervention or even exploitation.

Design fiction is one way to consider these possible futures. As a conflation of design, science fact, and science fiction, the medium is a method for exploring ideas, implementation strategies, and consequences [6]. Importantly, as Baumer points out in an introduction to a set of fictional conference abstracts, these visions of tomorrow can help shape the research directions of today [3]. Lindley further proposes design fiction as a methodology for considering the ethics of radical digital interventions [12]. Proposing our design fiction as an ethical provocation and a starting point for conceptualizing complex problems ahead in our socio-technical future, we ask: how could a vision of tomorrow inform the ethical considerations of the research we are conducting today? Where is the line between research and privacy, utilizing data insights and manipulation?

Written as an API, the piece situates itself both in technical and social literatures of computing. Questions have already been raised about the ethics of corporate experimentation and the fine line between product testing and harmful intervention [13]. Research has shown that users may not really understand what they are consenting to when agreeing to a terms of service [2,10]. They may also find certain uses of their data to be ?creepy? or invasive when it comes to behavioral advertising [19]. When asked about the process of data merging and aggregation, users tend to feel they are not the ones receiving a true benefit [5].

Though these user attitudes may raise red flags, research and industry continue expanding our capabilities in this area. In computer vision, deep neural nets have been a boon for new

models that aid in extracting emotion from facial images posted online [4,11]. Text is no different as research continues to improve our ability extract emotional insights from syntax tokens [1,14]. Separately, researchers have proven capabilities to make mental health inferences using social media data [7,8]. Typically, future directions for this kind of work involve technology design for helping people. However, there are other potential uses for this technology, including online marketing tactics.

If we consider the bleeding edge of marketing and artificial intelligence, we see very similar forms of emotional targeting being brandished as the next wave [16]. Yet, when users actually find out how they are being classified on psychological and emotional terms, it foments anger and is seen as ?overstepping boundaries? [17]. In academic circles, researchers such as Zeynep Tufecki and Kate Crawford have stoked debate around new kinds of privacy harms caused by advancements in AI and algorithmic methods [9,18]. Their concern is based on the fact that predictive inference is now able to go beyond what users openly disclose about themselves.

Ad Empathy and its API Resource offer a demonstrable grey area in technology ethics. The product very clearly meets the path we are trending toward, yet it should provoke some sense of caution or discomfort in its ability to find users at their most vulnerable moments. Without a doubt, this kind of system will become possible and machines will continue pushing the limits of our cognitive capacity to recognize manipulation, presenting ethical issues that are worthy of close consideration and skepticism. As a discussion piece, the Ad Empathy design fiction should work to ground debates around fair use of data, and the boundaries of ethical design. We hope Ad Empathy offers a point of negotiation around how to move forward relative to this plausible future.

4.2 Ethics Education in Context: A Case Study of Novel Ethics Activities for the CS Classroom

4.2.1 Prologue

This section looks at a novel curriculum I created for teaching ethics in the CS classroom. Originally designed for a human-centered computing class, the goal was to infuse ethical content *in situ* while teaching technical engineering skills. This class leveraged a number of novel activities to promote ethical thinking about the CS discipline. Why it is being added to this dissertation is because several of the activities and homework assignments explicitly involved narrative. Calling these out ahead, in our class we used narrative three ways: 1) to explore the possibility of future negative consequences for technologies they were designing for their class project; 2) to provide tangible examples for complex ethical problems discussed in class; and 3) to gain an appreciation of how difficult it is to predict the future.

What this involved in the course was several narrative-centric activities. We watched video clips and discussed stories to get at tough ethical problems that were better explained through grounded examples. For one assignment, students read short stories written by friends of mine that each explored possible futures where technology led to unforeseen consequences. The students were asked to analyze the stories in terms of their feasibility and expression of human-computer interactions. Students were also asked to explore their own software projects using narrative to try and understand the best and worst case scenarios. These narrative exercises were highly successful, as will be seen in the following essay. Several students openly called out these exercises in their weekly journals as some of the most impactful on their thinking.

4.2.2 Introduction

There is no disputing that computer scientists should be trained in ethical thinking alongside developing their technical skills. The ACM, IEEE, and ABET have all emphasized a need to prepare students to think and act responsibly, while grasping the applicable legal and business challenges

related to their practice. We are now seeing the importance of this training, as decisions made by computer scientists increasingly shape our public and private lives.

Examples such as racial bias found in risk assessment systems [27], severe privacy violations occurring internally at Uber [48], the expansion of filter bubbles and propagation of fake news [123], the proven difficulty of robust anonymization [238], and the rapid advancements of predictive inference using Big Data [182] suggest that computer scientists are becoming some of the most powerful moral agents in today's world. However, traditional ethics education for computer scientists may not include practical and timely training on how to weigh the consequences of their decisions. We therefore suggest that it is critical to incorporate ethics education as a continuous and practical thread within CS curricula.

In this paper, we offer our experience adapting an upper-level undergraduate Human-Centered-Computing (HCC) course to stress ethical thinking throughout the process of learning the fundamentals of human-centered design and evaluation. Our goal was to expand the current repertoire of *in situ* learning activities that require ethical judgment and to evaluate students' reactions to an infused ethics and engineering practice course. During the process, we built on prior work that has used project-based learning [184] and current events [30] to motivate realistic ethical problems while further piloting several new activities.

We found that students responded well and were even excited by having to apply what they learned to complex ethical situations. Many of the activities we piloted show promise for being adapted into other courses such as machine learning, data science, software engineering, and algorithms. Here we discuss existing models of ethics education in CS, the structure and components of our course, examples from student assignments, and results of a pre/post-survey. Finally, we unpack these results to make concrete suggestions for how other educators could reuse our material in other CS education contexts.

4.2.3 Literature Review

ABET's accreditation standards [9] and ACM's Code of Ethics [24] lay the foundation for ethics education in computer science. In light of these top-down guidelines, many efforts for ethics education in computer science have cited them in their approaches [68, 265, 334, 184, 104, 78]. There has been further influence from the ACM/IEEE Joint Task Force on Computer Engineering Curricula to adopt courses that build professional experience including dimensions of law and business practice [8].

ABET mandates at least 20 credit hours in professionalism–including ethics along with business, social impacts, teamwork, communication, design, and law [9]. This has given rise to the creation of many one-off ethics courses [104] or project-based courses that integrate concepts of business and law into the material [184]. In the ethics education literature, one can find successful project-based models such as Purewal, et al's course that included a service-learning component addressing e-waste and sustainability [265].

Supporting the creation of a CS ethics course, researchers have argued for classes that incorporate discussion of ethical dilemmas, designing rubrics to aid in the evaluation of such a course [284]. In fact, courses covering social impacts of computing have proven ability to increase interest in CS degrees [104]. Other educators have pointed out the potential for courses that integrate the rich material available in current events, multimedia and film, and short essays that cover topics across philosophy, privacy and civil society, intellectual property, AI, whistle-blowing, security, hacking, piracy, etc [30].

Some schools offload the burden of putting together an ethics course to their philosophy or social science departments. Though, teaching ethics outside of a technical context often leaves students with the impression that the material is irrelevant to them [104, 302]. This problem manifests as a general theme in the literature on CS ethics education—that isolating ethics into a separate or external course makes it appear as a side issue to computing [78].

Our work is an attempt to address this gap between technical material and ethical consid-

erations in the CS curriculum. With a clear desire in the field for improved ethics education [239] and many documented ideas for how to design interesting material, we hope to contribute to this literature by reviewing our own attempt to design a class that integrated core CS content and ethics.

4.2.4 Course Overview

The course was an intensive five-week implementation of a undergraduate-level human-centered computing foundations course. The class met 3 times a week for 2.5 hours per session. At its core, the class taught methods for prototyping and evaluating computing systems from a user-centered vantage. Our course further emphasized ways in which the design, development, and deployment of technologies have human consequences. Importantly, the curriculum helped students assess and plan for those consequences and hone the skills necessary to be socially-conscious and responsible engineers. Throughout the course students were assigned weekly reading reflections (10% of total grade), individual assignments (30%), milestones for a course-long group project (50%), in-class workshops (ungraded), and a participation requirement (10%).

4.2.4.1 Participants

The course consisted of 31 students (8 female; 23 male) that were primarily computer science majors (23). Beyond CS majors, the class also had one masters student and seven students seeking computer science minors or certificates.

4.2.4.2 In-Class Activities

Class time was split up into two parts: lecture and workshop. The lecture was 1-1.5 hours and was given by the primary instructor or an invited guest lecturer. Workshop time consisted of active learning experiences that synthesized lecture content into a practical exercise and discussion. Often, the active learning exercises would allow students time to process dimensions of their group project in relation to course content.

Throughout the course, we brought in a total of five guest lecturers: a researcher focused on co-operative ownership of data and software, a privacy lawyer with practical knowledge of the laws and regulations relevant to computer scientists, an artist specialized in typography and layout, a researcher focused on terms of service agreements and online harassment, and an emeritus professor who detailed his experience watching a 50-year transition of technology. These lectures allowed the class to gain perspective in legal, business, psychological, and historical dimensions of technology—all of which aided in presenting a broad, robust conception of social impacts.

Workshops were primarily done in small groups or within students' project teams. Students were often asked to brainstorm and share ideas with one another with the expectation that disagreement and debate could occur. In order to introduce students to this kind of dialogic atmosphere, the workshop for the first class was a "spectogram" exercise. We had all the students stand in a line and asked them a question formed along two polar extremes. Students then had the opportunity to physically stand anywhere along a spectrum between the poles that best represented how they felt. We then asked a few students from different places on the spectrum why they took this position and then allow students to reorient if a particular point changed someone's mind. We did this exercise with three questions, respectively: 1) Do you believe facebook is good or bad for society?; 2) Do you believe face recognition technology is good or bad for society?; 3) Do you think it is good or bad for Facebook to use face recognition technology to identify the faces of untagged people?

In week 2, we had a workshop that asked students to try and identify the primary stakeholders that would make up the user communities for the systems being developed in their projects. While drawing stakeholder diagrams, students were asked to consider the different "personas" [263] relevant to who may use their app and further identify competing incentives that may align or conflict between these groups.

As the course developed, these workshops became more sophisticated, requiring a deeper understanding of technical and social concepts. Week 3 included a workshop where students were asked to design "capture" systems [19] and come up with metrics for how they would extract insights out of particular data sets. They were first shown an example of the kind of data collected by our

university's online courseware platform. Given that dataset, students were asked to come up with a metric that would allow someone to identify good students to recruit for graduate school. Next, they were asked to design a metric that could be used to analyze Facebook profile data and give a user a "wealth index." Finally, they were asked to design a capture system for a news aggregator that allowed publishers to upload articles and metadata. Their goal was to be able to accurately flag fake news upon upload. At the end of each task we did a talk back where we would identify biases, unfair consequences, or gameable aspects of their designs to help solidify notions of data fairness.

In the classes where a guest lecture was invited, the workshop would attempt to link new ideas from the workshop into core course material already being applied to their projects. For instance, when a privacy lawyer came to lecture about the EU's upcoming General Data Protection Regulation (GDPR) implementation, students were asked to then reconsider their UI designs under the constraints of this regulation. They were given an hour to review at least one of their UIs that contain user data and search for compliance issues and ultimately consider a redesign that would solve these problems. When a terms of service and harassment lecture was given students were asked to consider the social norms that would be important for proper use in their project designs. The workshop then gave them time to consider bullet points for a harassment policy or terms of service and what design decisions these changes may affect in their implementation.

4.2.4.3 Individual Assignments

Outside of class students had three types of individual assignments: weekly reading reflections, participation submissions, and an individual applied exercise regarding a topic from the week. Each week had 3-4 required readings and to complete the reflection students had a choice of 3 questions to answer. They could either choose one question and dive deep, writing at least a full single-spaced page, or choose two questions and answer them in a summary form writing at least a half page. The questions were written to force students to recall an aspect of the author's argument or main point and then critically reflect on a dilemma this presents engineers.

Participation submissions were simple responses to fulfill weekly point requirements that could otherwise be attained by speaking in class. Two points per week were required to receive full participation credit and students who did not directly respond to a question or discussion in class could alternatively choose a topic or discussion from class and submit a short write-up. This either detailed their opinions on the topic or offered an example from the media or research that availed deeper complexity.

The solo assignments were meant to give students a chance to apply skills learned in the class outside of the project. One assignment included piloting an observation protocol by watching someone do a common task and interviewing them about how they performed it. Another allowed students to implement a design experiment, applying ideas taught in our lecture on design aesthetics, asking students to play with alternative layouts, colors, and interactive elements that could be incorporated into a website. An ethics exercise we piloted involved students reading a sci-fi piece and discussing the underlying technology, the assumptions about how society was changed by this technology, and ultimately analyzing the likelihood of the story's events happening in real life.

4.2.4.4 Team Project

The anchoring component for students was the course-long group project. Milestones involved prototyping the interfaces and functionality for a computing system, evaluating their design, modifying their design based on evaluation, and reflecting on the social impacts of their system if it were to become a commercialized product. Week 1 kicked off with a simple team statement that explained who was working together and how they would submit their assignments (e.g., pencil/paper drawings, digital mockups, CSS/JS code). Week 2 required teams to come up with a problem statement for their project and an initial design concept that included visual mock-ups to describe their initial approach.

The project continued with the students having to apply course content to its development. Week 3 called for a cognitive walkthrough and think aloud study to be performed on their first interfaces. Week 4 asked students to respond to their findings, showing a revised version of all

their interfaces. Week 5 finally asked for a report that explained the evolution of their design, a UI showcase that explained their system's workflow, and detailed research into several dimensions of social impacts that should be considered if this project was implemented into a full system.

4.2.5 Results

An important premise of our effort is to introduce ethics components *in situ* without sacrificing the core contents. Since our course was not designed as a controlled experiment, it is difficult to draw a definite answer whether we succeeded in meeting this premise. However, judging from the projects student groups had delivered, we did not notice significant difference in quality from previous offerings of the same course. Nor did we notice significant difference in grade distribution compared to previous offerings. In a post-survey, when asked to specify main takeaways from the course, many had mentioned core HCI concepts such as "user testing methods", "considerations for design decisions", and "heuristic evaluation." Taken together, these indicators suggest students were indeed learning the core contents effectively. In the rest of this section, we turn our attention to evidence we uncovered regarding ethics learning.

4.2.5.1 Student Responses and Reflection

Looking to individual assignments provided a viewpoint into how the students were processing the course material and ultimately what questions and considerations stuck with them. On a weekly basis our instructional staff read all assignments, pulling out general themes and looking for interesting comments that provided a launchpad for discussion at the beginning of class. In some instances, we would highlight two perspectives that had opposing considerations or solutions to spur discussion around trade-offs and differing assumptions.

Here we offer a summary view into some of the interesting questions and considerations raised by students that signaled a deeper engagement with ethical thinking. Starting with week one, an encouraging result were reflections that portrayed engagement with the "spectogram" exercise, as reflected by a comment "**The topic on the first day of class: facial recognition, was**

particularly interesting to me. While I thought I had a right stance at the beginning, every body's other ideas made me keep think about what I had thought already differently."

A strong sign that ethical and critical thinking is occurring is seeing a student weighing the trade-offs of a particular dilemma. For example, the dilemma of "manipulation vs personalization" was a common topic when we had students read an article by Tristan Harris called, How Technology is Hijacking Your Mind [160] that discusses the cognitive tricks designers can play on users to maximize on-screen time. A student wrote "**Some individuals have more addictive personalities and the cognitive load needed to stop enjoying the constantly refreshed content is much larger. In this way the technology [to manipulate users] seems unfair...but you can change your settings, you can turn off your phone... I can see the argument for why it is fair.**"

When reflecting on ProPublica's article, Machine Bias [27], another student exemplified a strong understanding of how the concept of fairness may differ between a user and an engineer. This student wrote "**In a system where the people being evaluated are given no choice and are having their lives changed based on the results, fairness is of the utmost importance...As we have learned in class, someone who knows a system well because they built it is unlikely to be able to replicate using it in the same way someone who doesn't know anything about it would.**"

We also found that students were applying ethical reasoning in their personal life and professional work. For example, a student talked about encompassing skills from class in an internship. The student first commented on their past viewpoint before the class "**[Before] I would only think about the people that actually used the service or product, instead of also including who can also be affected.**" and then reported how the class had changed their viewpoint at work: "**[The class] made me think pretty hard about the user-communities of the start-up that I'm volunteering on, and how thinking of the user-communities for our product will greatly increase the usability and make sure that all the communities**

(not just the immediate user) are fairly treated."

In the same reflection, this student commented on a guest lecture detailing the differences between privacy laws in the United States vs. the upcoming European Commission's GDPRs and how they could affect the start-up for which he is working. Through all these quotes we see a rich engagement with ethical questions as they relate to technical and professional thinking. It should be noted this was only the tip of the iceberg and we had some students go deep into this thought process; one even wrote a 5-page lyrical poem about the ethical dilemmas around search engines.

4.2.5.2 Pre/Post Survey Results

We collected pre- and post- class survey responses from 30 students in the class. The surveys contained short-answer, open-ended questions related to the ethical implications of technology and takeaways from the course. The responses to these questions were coded by two research assistants. The responses were given a score of 1 if it contained any mention of ethics, social implications, privacy, or consent. The coders read each of the responses independently and assigned a rating. Differences in the responses were resolved by discussion.

Findings from these survey responses further suggest that embedding ethics into the curriculum increased the amount of students who believe ethics are important to their careers. There was a significant change in the post-survey responses: more students considered ethical implications in their statements. When asked "What questions would you ask to a potential employer if you were being offered a job to design an app that sorts and displays people's photos across all their media platforms?", 30% more students (Pre:10 out of 30; Post:19 out of 30) considered ethics or ethics related concerns in their post survey responses, such as, "**How would you like users to authenticate other devices? What sort algorithm would you like to use? What functions should the program have? What feelings should the user have while using this product? How are photos being stored and who has access to them?**" Additionally, when asked "If you were to make an app where networks of users could share geolocations for house parties—what users might have issues or who may this harm?", 33% more students (Pre:9 out of

30; Post:19 out of 30) mentioned privacy and social impact problems. For example, "**This could harm anyone who does not want their location known, or people who have their house registered as having a party incorrectly, as well as anyone who was having a party, but didn't want it broadcasted.**" The responses showed that our course with ethics interventions throughout had a doubling effect on students considering ethics in technology design.

We included three extra, open-ended questions in the post-class survey. We categorized a response as integrating ethical concepts if the student explicitly mentioned ethics, morals, or privacy, or if they implied ethical decision making through thoughtful contemplation of the social implications of their work, potential harm to others (i.e. empathy, identifying different stakeholders), or caution when approaching and designing a technology solution. It should be noted we did not prompt or otherwise guide students to discuss ethics or social impact for these questions.

When asked, "What will you take away from this course?" 26 out of 30 student responses included ethical considerations, such as, "**I realized that technology can have a much larger impact than I initially thought. I never really took the time to analyze possible future implications or think about privacy. I also never really thought about evaluations and using those to enhance whatever you're working on.**" This shows both a greater perspective of how technology impacts society as well as an appreciation for practical design methods, but other students went further in expressing the need for ethics to be baked into all aspects of technology development. One such student wrote "**There are ethical considerations in almost all aspects of design and implementation, something I hadn't really thought about previously.**"

In the next question, "What's the one thing you've learned in this course that will be most applicable to your career?" 14 out of 30 student responses included ethical considerations. Although this is lower than we had hoped for, we believe responses to this and other questions such as the one before show students' willingness to consider ethical implications in their work, especially when it came to privacy, as answered by one student: "**I'll take more time to consider the implications of privacy policy in my startup ideas.**"

Furthermore, in the final question, "What is one thing you'll do differently following this

course?" 25 out of 30 students said that they think about ethics and the social implications of work differently, including their future work. One student commented "**This course has taught me how to consider my ideas from an all encompassing standpoint, from design, to ethics, to law, etc, and i'll continue to use this frame of mind, especially in my senior project this coming year.**"

Another objective we had with ethics was to explore the tensions that arise in software companies when business goals are introduced. Students were able to articulate such tensions, as one student noted "**That everything needs a way to make money, and often that means sacrificing certain ideals like "free no matter what," privacy, or how you keep people using your software.**"

Finally, we sought to highlight the importance of realizing that humans are naturally bad at predicting what the future holds. This is critical because when developing a technology, such as a classifying algorithm, it is difficult to know how it will be applied in society. And this was expressed in student responses as well, such as "**Predicting the future is next to impossible, so it is important to pay attention to current trends and not get stuck in an obsolete 20 year plan. New laws such as the GDPR mean that as a designer it is important to ensure your business model is not dependent on practices that would be made illegal. Always important to consider values of users and compromise accordingly.**"

4.2.6 Discussion

The results of this course were encouraging. The use of current events, real-world problems, and artistic provocations, even when reduced in complexity for pedagogical purposes, amplified student engagement. Many students brought up the fact that this was the most thought-provoking course they had taken and that it opened their eyes to new dimensions of their field. Our course is a foundation course for HCI. We were not forced to sacrifice content nor difficulty by adding these ethical and social elements. In fact, exercises such as designing capture systems and metrics to understand data bias, were almost too technically complex for an undergraduate course. The biggest

limitation of our course was over-motivating some of the social issues and being unable to fully address their technical counterparts, such as diving deep into machine learning or cryptography, due to time and content scoping constraints. However, when students later do take topic classes on machine learning or cryptography, it would be ideal for them to receive another dose of ethics training.

Encouraged by the results, we adopt a position on CS ethics education that includes the following principles:

Continuous Ethics education should happen in small doses throughout the curriculum rather than in a one-off course.

In Situ Adding an ethics component to a class assignment relevant to core course content shows that ethics and engineering thinking go hand in hand. This would imply discussing bias and anonymization in a data science or machine learning course, privacy in a computer vision course, etc rather than treating ethics as a set of concepts to learn away from technical material.

Perspectival CS We ought to turn our value equation toward students' abilities to identify multiple perspectives about computing issues (i.e., recognize a dilemma) and translate that dilemma into competing technical choices.

We see many ways these principles could be adopted. A centralized, but simple approach could involve the instructors of a CS department coordinating to add small reflective writing assignments to several CS course in the program (e.g., A Computer Vision Course with a successful reflective component [91]). Though not all courses might lend themselves to as much ethical context as HCC, many technical courses (AI, machine learning, etc.) could integrate ethical components. Some of the exercises we did in our class could be easily adapted. Our exercise in defining a metric on a dataset to discuss bias and trade-offs would be perfect for a data science, algorithms, or machine learning course. Testing out the enforceability of different harassment policies could be

done in a social computing class. Evaluating different stakeholder incentives to balance privacy perspectives would be useful in a cybersecurity class. Learning about how one may mislead an audience with data would be perfect for an information visualization class.

4.2.7 Concluding Remarks

Discovering novel and engaging methods for training responsible engineers that do not sacrifice learning technical skills will continue to a central problem for CS curriculum design. As our case study shows, infusing ethical dilemmas and social challenges in the curriculum is by no means a crutch for a course, but can amplify interest. We hope to continue by expanding on the activities we piloted in this case study and encourage other education researchers to do the same. It is our position that until ethics education is threaded throughout a student's development of their computing skills that we will continue to see ethics treated as a side issue rather than a central asset to CS.

4.3 Quantified Self: An Interdisciplinary Immersive Theater Project Supporting a Collaborative Learning Environment for CS Ethics

4.3.1 Prologue

This section reflects on the educational effects of working on an immersive theater piece about the ethics of data-driven technologies. Later in the thesis a more robust depiction of Quantified Self, an immersive theater piece that invited audiences to explore ethical issues around data, will be discussed. Here the production team is interviewed to try and understand the educational affordances provided by doing cross-disciplinary projects. In this piece the script and production acted as a **template** for the entire cast and crew to use for thinking through ethical issues. As this section will avail, this turned out to be a very useful way to engage students from multiple disciplines coming in with varied interest in this process. Students across the board walked away with a heightened awareness of the issues, many forming strong opinions and changing their online behaviors as a result.

4.3.2 Introduction

Education around computing is highly valued and even non-technical students grow strong interests in developing computer literacy throughout college and career. Despite the ubiquitous impact of computing in their lives, without becoming a computer science (CS) major, it is unlikely non-technical students have the opportunity to formally study privacy, data ethics, or socio-technical problems. Within CS departments, on the other hand, we have requirements [9] that our majors leave with an understanding of computing ethics and social impacts. Even with most departments having a course dedicated to social impacts, CS students rarely gain perspective on how outsiders conceptualize their work prior to entering industry. We see this as opportunity in the CS curriculum to connect CS students' study of social impacts while offering non-technical students the chance to develop knowledge and perspective on computing.

As an approach to bridging this gap, we devised a novel, year-long project around creating a theater piece about data to offer an opportunity for engineers and non-engineers to work together and learn about crucial problems from one another. This performance, titled Quantified Self, was also meant to create space for dialogue among engineers and non-technical users in the audience. However, for this paper, we focus on the educational dimensions for those involved in the production of the show.

Interdisciplinary education is already being discussed as priority for the future of computing education [21]. At the same time, art and design fiction are being heralded by communities in education and human-computer interaction (HCI) as methodologies for collaboration [116], thinking about the future [28], and engagement between engineers and users [58, 248]. Thus, we saw an opportunity in the production of a highly technical art piece to structure a space for educating technical and non-technical students.

In this paper, we present our technical theater piece as a case study toward promoting technical art projects as a promising way to create interdisciplinary educational opportunities. After reviewing the motivating literature, we provide a full description of our project, highlighting

the choices our team made to elevate learning opportunities for all involved. We further offer insights from post-project interviews with the students involved in the production. Our findings indicate that technical students were given opportunities for learning technical skills while gaining insights into social impacts of their work. Further, non-technical students developed awareness, subject-matter interest, and formed more complex opinions about computing in society.

4.3.3 Prior Work and Motivation

4.3.3.1 Interdisciplinary CS Education

Computing is a pervasive function of any educational or career activity. Whether doing graphic design, making a website, writing an algorithm, or sending photos to a colleague, modern life requires some level of computing skills. Due to this expansion of the field, computing practitioners take multiple forms—from "creative coders" to "designers" to "back-end engineers." In light of these changes, researchers are discussing the need for interdisciplinary learning within CS education [21, 345]. Seeing the future need of integrated, interdisciplinary approaches to world problems, consortia such as the National Academy of Sciences are promoting interdisciplinary research and learning more broadly [286].

Adopting interdisciplinary practices in CS higher education has several challenges to address: a) getting CS students to develop skills beyond writing valid code; b) offering non-CS students access to education about computing; and c) structuring opportunities for technical and non-technical students to learn with and from one another. In terms of (a), ABET already has pushed the requirement for CS departments to integrate education in related outside areas such as law, business, and social impacts of technology [9]. Researchers have successfully designed courses with service-learning components [265] or that motivate interest in CS by incorporating current events, media, and arts [30].

Towards (b), it is rare to find engineering departments taking on the responsibility of educating non-technical students. Rather, humanities departments tend to offer technology-focused

courses. When engineering departments do create such offerings, however, there are promising results such as courses about privacy meant for non-technical students [122]. There has also been progress in using artistic means, such as dance [252] or creative writing [177], to create opportunity to interest a more diverse body of students in CS.

Many case studies have emerged showcasing fascinating efforts for (c). Work between Carnegie Mellon and Disney Research shows that combining efforts between artists and engineers led to novel projects and technical applications that may have otherwise not been feasible [168]. Hybrid classes between CS and journalism [259] or CS and bioinformatics [133] have been successful at getting computing skills out to other fields and giving breadth to CS students.

A common theme for creating outside engagement in science is to incorporate arts. For practitioners, conferences such as Ars Electronica bring highly-skilled technicians and artists together to push boundaries of the discipline. In the classroom, technical and scientific classes that incorporate art components appear to work at all levels. From K-12 classrooms that get students into math using poetry [148] or computing using digital art [347] all the way to college classes making dancing robots [299] or professional artists and scientists co-creating theater [116]. Throughout these examples we find higher levels of engagements, reflection on socio-technical issues, and refreshing new approaches to computing.

4.3.3.2 Design Fiction / Deeper Engagement

Success of arts and science collaboration goes far beyond interdisciplinary classes. Within the realm of HCI, design fiction has blossomed as a promising approach for thinking about the future, discussing ethical implications, and engaging a broader public [307]. Design fictions allow technical work to happen in a space that blends fact and fiction and welcomes modes of critique and social thinking [58].

Long before it was an HCI trend, sci-fi operated as an intersectional space for scientists, hobbyists, creators, and the public to conceptualize the future and what technology means to human enterprises [58, 188]. Many artists and designers who want to pose questions about how

technology is impacting people employ "speculative design" as a method for asking questions and generating dialogue between communities and stakeholders [28]. HCI researchers have adopted a framework of "enactments" using speculation and theatrics as a methodology for users to experience possible future [124, 248]. In turn, gaining insights about public perceptions and attitudes toward technological change.

Seeing these successes in both interdisciplinary education through art and design fiction for public engagement, we began considering a creative endeavour that connected these two potentials. Our first project involved obtaining a small internal grant from our engineering school to support making an art installation as the class project for an upper-division data science course. This pilot, involving 30 students working in groups to make 6 art pieces about data, showed us promise of expanding. Not only were students enthusiastic to think through social dimensions of data applications, but the prospect of a public unveiling of their work amplified their desire to take on challenging technical work.

Following this, we began laying the foundations for a more robust project that would bring together students from different departments to create, converse, and ultimately learn from each other. For the remainder of the paper we report on the resulting year-long project, QSelf, which brought together students from seven departments to create an immersive theater performance about data ethics. Our project offers a case study relevant to the literature of interdisciplinary CS education, collaborative learning, and CS ethics education and provides a model appropriate for adoption as a capstone project within a CS curriculum.

4.3.4 Approach

QSelf was a student-led initiative to create an immersive theater production that brought together a sci-fi drama and interactive art installations to create an explorable world for audiences to think about the future of data in our society. Audience members were able to connect their social media accounts at ticketing to offer their own data to personalize the experience. From the start of the project in autumn 2015, we aimed to make the production a collaborative learning opportunity

to bring together students from diverse backgrounds. Students involved in the production were offered multiple modes for engaging with new material and each other. Students were able to take on roles as actors, production staff, technical staff, or scenic designers. Here we detail the full scope of the project.

4.3.4.1 Team and Project Structure

The production team was mainly composed of students. The project lead was a third year PhD student in computer science. While a number of university professors and industry experts served as advisors, almost all the decisions and creative outputs were made by the students themselves. Two other PhD students and a professional data scientist served as co-producers. The rest of the team were 19 students - 15 undergraduate, 3 masters and 1 PhD. Students came from 7 different departments crossing technical and non-technical majors: 5 from computer science, 1 from electrical engineering, 9 from theater/dance, 1 in studio art, 1 from music, 1 in neuroscience, 1 in english, and 3 in an interdisciplinary technology department. The production crew was split up into two teams: 1) a technical team that designed scenic elements and engineered the interactive exhibits and 2) a theatrics team dedicated to preparing and producing the performance.

While this project took place outside of a typical class structure, we found ways to incorporate the production into university requirements for students. Theater and dance majors have to spend a certain number of hours supporting theatrical productions which this was able to count toward. Most of the computer science students took independent study credits under the supervision of our project lead. One undergraduate CS student became employed as a research assistant working between the data scientist and project lead. Some students were employed by the project as an intern while others volunteered out of interest.

4.3.4.2 Pre-Production

Narrative

The script was written by our project lead prior to the start of the project. It set the

basis for 8 characters portrayed by 6 current students and 2 recently graduated students. The characters represented different perspectives on technology: a corporate CEO, a hacker, a journalist, a law enforcement agent, a data scientist, a psychologist, a marketing strategist, and an AI-driven android.

Laden with ethical issues, the script was designed to create an open structure where conversation and technical interaction could occur. We addressed five primary ethical issues within the story line: 1) implications of privacy policies; 2) the psychological effects of data presentation; 3) the use of personal data to infer information about a user; 4) the effects of ubiquitous personalization; and 5) the use of personal data as a commodity. These issues were represented by struggles between characters within the narrative and grounded by the interactive technical exhibits. The play was presented in 4 acts: 2 where scripted performance was being witnessed by the audience and 2 where the audience was freely exploring and the cast was improvising their roles while interacting.

Our goal was two-fold. On the one hand, the arts students would have the chance to understand more complex dimensions of technical issues by having to articulate and embody their roles. On the other hand, the technical students would have to learn how to translate these ideas to the non-technical cast while taking on the technical challenges presented by working with real data in the exhibits.

Exhibits

While fitting the themes of the narrative, the exhibits were finalized in terms of functionality and appearance through a dialogue between the whole team. Everyone on the technical team had an open period to propose the design of an exhibit which would be vetted and revised by the project leadership. Knowing the thematic constraints, technical students were given time to do research and learn the computing tools necessary to build the exhibits.

Our goal was to make the creation of the exhibits be a process where technical students could learn new skills while ensuring a dialogue between the technical and non-technical students occurred in order to make the ultimate adoption of an exhibit into the production as coherent as possible. For most of the students this was the first time they applied their classroom skills to real

data for true users. Beyond small prototypes, they were given the chance to learn about public APIs, data processing pipelines, and front-end libraries for presentation. Throughout the process we diligently discussed the reality of privacy expectations from the audience members. Our data scientist designed an encryption protocol that the students had to use to keep data secure once we ran our systems on live data. From conception to implementation each exhibit was discussed at our weekly technical team meetings both in terms of how it represented our thematic issues and its technical specifications. This meant students gained technical and social perspectives in parallel.

One student designed a non-technical exhibit that represented perspectives critical of how technology has affected human relationships. This exhibit was designed in closer collaboration with the theatrical team as it required participation from the audience each night.

Training and Rehearsals The rehearsal process was led by the theatrical team, though the training of the actors involved collaboration with our technical team. Multiple meetings occurred where the technologists trained the actors on how to work with the exhibits. Certain actors were trained on particular exhibits related to their character. Not only did the actors get trained on how to use the technology, but also they gave feedback on what a non-technical perspective was so that the tech team could tweak the exhibits.

One of the more interesting aspects of the work involved training the actors to be able to speak as technologists. There were 5 (out of 22) rehearsals that were dedicated to working with the actors on their articulation of the tech concepts and testing interactions with the exhibits. Starting with the first reading of the script, we brought the entire team together to hear the full performance. Throughout we paused and explained vocabulary to the actors and offered them auxiliary terminology that could be useful for improvisation. Given that it was an immersive theater piece, the actors had to be prepared to speak impromptu to audience members coherently about technology. During dress rehearsals the actors had the chance to practice their improvisation with our technical team and then afterwards we did feedback sessions. It should also be noted that one actor actually was a CS major and he was crucial for giving continuous feedback to the theatrical team outside of these dedicated sessions.

4.3.4.3 During the Runs

Participation during the runs of the show were mandatory for all students involved in the production. Of course, actors were the primary focus while technology students staffed the show as tech support and assistant stage managers while costumed as corporate employees. This further gave the technology students a chance to listen and interact with the audience during the interactive parts of the show.

Within the show, each audience member had a wrist band that unlocked the exhibits and unencrypted their data. Cast members often got to use the exhibits in tandem with the audience and the tech team was able to observe the reactions to their exhibits first hand. After the performance each night, we held talk-back sessions where the production team and actors discussed elements of the show with audience members. Topics ranged from what it was like to play different characters to ethical questions about the script. We further used this as a chance for the audience to talk among themselves about the issues raised in the show.

4.3.4.4 Post-Production

After the production, we elicited feedback from all production team members about their experience preparing and producing QSelf. Students were asked to reflect within individual interviews upon their learning, challenges, and recommendations for the future run of the production. This final reflection was individual to allow each person to open up about what worked or did not for them. In the next section, we report the findings regarding educational opportunities the show afforded. However, regarding collaborative process, we found a general interest for more collaboration between the actors and the technologists earlier and more frequently.

Another major interest was for the technical exhibits to be pinned down earlier to allow the cast to get more comfortable and knowledgeable. Similarly, the technical team appreciated the ability to be inventive, but would have preferred the technical challenges be structured more top-down to allow them to develop skills in a more constrained environment.

4.3.5 Evaluation

As a show, QSelf was a success, selling out all 6 performances and bringing in over 240 people with many more on the wait list. As a collaborative learning experience, which is the emphasis of this paper, we found strong evidence of learning among the students who worked on the production, the type of learning that could not have occurred in traditional classroom settings. Below we present two sources of evidence: the exhibits designed by the students and post interviews.

4.3.5.1 Exhibits

Students in our technical team were tasked to design and develop novel interactive exhibits that tied to the performance. In total, 10 exhibits were built. To highlight a few, one exhibit was a 4-player game where the players' Facebook and Twitter posts were shown anonymously and the game was to see if people would own up to what they said. Another exhibit was a magic mirror where a player's private data was used to display personalized greeting messages in a private room on the set. The development of these exhibits presented various degrees of technical as well as creative challenges. Students were able to overcome these challenges, communicate across diverse teams that included designers and non-technical users, adopted strong version management skills using GitHub, and got first-hand experience of the stresses involved in preparing for a software launch. In terms of CS skills, all our front-end systems were built in either ReactJS or P5.js, which are popular JavaScript-based UI frameworks widely used in the industry. The back-end server was built in Python to securely manage the social data provided by the audience members. The students not only had learned and honed their Python and JavaScript programming skills but also had become familiar with data skills such as authenticating users (OAuth), retrieving data from APIs, and visualizing the data. The skills they learned were central to those expected by CS majors. Moreover, they all made this project an experience highlight on their CVs.

However, the pressure of finalizing the exhibits prior to our first show turned out to be a challenge. One student commented that he was interested in the conceptual material, such as

privacy, but it got left behind because he spent too much time on debugging.

4.3.5.2 Post interviews

After the production, we conducted 15 interviews (5 from the tech team and 10 from the theater team), which represented 79% of the crew. Each interview took about 20 minutes. Our questions focused on self-reported learning, change in opinions, and technical content understanding. Each interview was recorded, transcribed, and coded. We looked for insight language, changed assumption/viewpoint, new information/knowledge, and behavioral changes. Overall, we found that participating in QSelf had impacts on students' technical knowledge, ethical viewpoints, and daily activities, especially related to the interdisciplinary opportunities made available by the project.

Participants expressed having insights through the production. When asked whether the experiences had made them more knowledgeable about the issues, most (13/15) were affirmative. The majority felt strongly so (8/15). "**Yeah. Absolutely! (P5)**". There is evidence that the experience helped students dispel certain previously held misconceptions. "**I didn't think [companies offering free services] would use my data for malicious intent. (P5)**" "**It opened my eyes to the fact that [my data] is being used for things, and that it's being bought and sold, and I had no idea that that was a thing before. (P9)**" This suggests that students did gain new knowledge.

However, there were limitations to the knowledge gained. A number of art students made specific statements that they became more knowledgeable only about the issues of data use but not about the underlying technical skills such as coding (P4) or data analysis (P2), even though that was of interest to them. Certain misconceptions were still held by some students. "[**Texting lets you] communicate with people easily without having to put all your information out there (P8)**", when in fact texting could still expose one's personal information. Two students (2/15) reported gaining no or minimal new knowledge. One explained that "**I already knew a lot (P14)**" prior to joining the crew. The other student explained that "**because I worked more on the front-end of the project, I wasn't doing a lot of the research tasks. (P11)**" But this same subject also reported that "[**the show**] **made me perhaps a little bit more skeptical (P11)**". This seems to

suggest while a student might not gain concrete knowledge, the student nevertheless had a change of attitude.

There were impacts on attitude and ethical standing. 12 out of 15 suggested changed assumption or view point. "**It's definitely made a change for me...paying attention to the fact that there's some kind of massive control here that I wasn't aware of. (P9)**" ; "**My opinions on how people view data sharing changed a bit, because people seem to be terrified, when people ask for their data, but have no problem putting all of their [data] on Facebook. (P15)**"; "**[Companies] are more monetarily motivated than I originally believed. (P5)**"

One noticeable difference between the technical and non-technical groups was that six art students made speculations about the future and how that could align with their ethical perspective; yet all of the tech students refrained from making speculations. "**I would love for us to take back our data, to understand, or to force some kind of understanding, that what we share still belongs to us and so it can't be bought and sold, and that shouldn't be a thing. (P9)**"; "**Eventually, I feel like legislation will get passed and at least reign it in. (P15)**"

Beyond knowledge and attitudes, many reported behavior modification in their daily life. 12 out of 15 made explicit comments on behavioral changes. Seven reported having changed certain behaviors as a result of the show: "**I'm more conscious of what's happening to my identity online. (P2)**"; "**I actually noticed that after Quantified Self, I just stopped posting on Facebook. (P13)**"; "**It did make me a little more aware of my presence on different sites and just being aware that everything is accessible. (P12)**" Those who reported no change were predominantly from the technical team.

11 of 15 provided personal anecdotes showing that they applied understanding from QSelf to their daily life. "**I thought it was just a happy coincidence that the shoes I was looking at...being advertised for me on this completely different website...[I felt] this is really creepy. (P9)**" ; "**I decided to Google myself...these links from this blog popped up, these photos [of my bags]...I have no idea where this came from. And that just made me a little more aware of (P12)**".

Several students expressed being more keen and equipped to have conversations about data sharing issues after the QSelf production experience. "**I had discussions with my friends and my roommates, ever since the show, about these topics and I think that they shared very similar views to what I had before the show. (P5)**" Moreover, the new gained knowledge allows this student to bring the discussions to a deeper level. "**And now afterwards, we can have realistic discussions about big data being a commodity. And how that shapes our society and how that shapes our perception and our interactions with everyone else in it. (P5)**"; "**It's something that I bring up more in conversations... people will randomly make a comment, like an advertisement being eerily close, and I can plug in and talk about Quantified Self, and people are always kind of surprised about how much they're sharing, without even realizing it. (P2)**" This suggests students have become teachers.

There was an appreciation among many interviewees of how the interdisciplinary nature of QSelf influenced their learning. There was evidence of mutual learning between groups. "**What came out of doing the project was knowing more about how tech people or computer scientists actually go about doing this. (P14)**" Learning also occurred during team meetings. "**It came up in one of our meetings, is that I only ever had to sign away my data rights once, for that to apply to all of the companies. (P5)**" Several members commented on the edifying experiences of playing a character interacting with the audience. "**[I] become more a little more articulate, especially 'cause I had to talk to the audience about it...and try to draw, elicit responses from other people so that makes you kind of think through it a little more. (P6)**"

4.3.6 Discussion

This project availed a lot of potential for the collaboration of technical and non-technical students for the creation of art, thereby learning critical technical issues related to big data. All students walked away with a better ability to articulate technical issues, many of them having changed their viewpoints or behaviors following the production. For technical students, there was a general takeaway that they learned about how the lay public perceives data technology. They

were shocked at how little the public understands of technology including how much data is really available through online services and how gullible they were to believe results from even simple algorithms. Having incorporated technical challenges into the project, this gained perspective came without skimping on technical work core to their major.

On the other hand, we learned that both dimensions could be improved. For technical students, going into the project with more structured technical work may have allowed for a better honing of skills. Further, offering even more chances for them to talk to non-technical cast may have allowed for deeper perspectives to emerge. For non-technical students, nearly all of them reported better awareness of technical issues and developed more informed and articulate opinions. For some it catalyzed an interest to learn more since few actually felt they understood the inner-workings of the technical systems. Despite these successes, holding short coding workshops would have improved the experience as many non-technical students stated regret for not learning more technical details. They too wished for more conversation with the technical team earlier in the process.

We believe with some minor improvements and coordination between departments, a project like this would make for a valuable interdisciplinary capstone project. The project was so attractive to students we were unable to meet the demand and bring in everyone who showed interest. It is very clear that on both ends students were eager to work with other departments. In the end students walked away with more than skills for their respective disciplines, but perspective and collaboration skills that will impact them in their careers and personal lives.

4.3.7 Conclusion

In this paper we presented an overview of an interdisciplinary theater project that structured a collaborative learning experience for computer engineers and artists/designers. The project was successful at building tech skills and gaining understanding of social impacts of technology for the CS students. The non-technical students became more aware and articulate of technical issues. Adopting a similar model, universities could make opportunities like this for interdisciplinary cap-

stone or course-long projects and give students a chance to interact with the public regarding their field.

4.4 Designing a Moral Compass for Computer Vision Using Speculative Analysis

4.4.1 Prologue

This section exposits my first attempt to devise a methodology for other engineering practitioners to question their own domains using speculative techniques and narrative. Written for a computer vision (CV) audience, the piece uses speculative scenarios to develop a basis for exploring possible futures of new CV technologies. The scenarios are interrogated from the lenses of technical feasibility, law and policy, and morality. Borrowing tools from future studies and risk perception literatures, I attempt to systematize my findings and prioritize the scenarios in terms of their likelihood and severity. Once I go through this analysis, I write longer narrative case studies on the issues found to be most concerning. These narrative case studies establish **targets** for debating ethical questions between practitioners or discussion in the CS classroom.

4.4.2 Introduction

Computer vision (CV) techniques are at the epicenter of excitement and progress related to recent developments in deep learning; specifically, convolutional neural networks [196, 162, 163, 357]. Concomitant with a surge in the success of machine learning systems is unprecedented access to new datasets [64, 23, 87, 108]. There is no end in sight for this growth in promise and applications. The massive availability of image data coming from commercial sources such as Flickr, Instagram, and Facebook, and the dispersed use of ubiquitous and smart camera systems has locked us into a future where our living image will constantly be monitored, captured, processed, and used to generate inferences.

Without a doubt, combining advanced machine learning with troves of image data is likely to aid human causes such as health monitoring [198] and accelerate efficiencies in areas such as

archiving [83] and traffic analysis [362]. However, with the blinding light of promise glistening, we must be careful not to miss that there are consequences and dangers to allowing these applications to run amok.

Over the past few years, we have seen many red flags waved that should caution researchers to how deep learning and CV may go wrong. Machine learning techniques have been critiqued for their ability to inherit bias and create discriminatory results on tasks that may have chilling consequences [27]. MIT researcher Joy Buolamwini began the Algorithmic Justice League after discovering that a common face recognition software failed to work on black faces [72]. Tech giant, Google, was forced to dial back image captions as their software regularly identified photos of black people as "Gorillas" [5]. Further, the threat of unwarranted or unfair surveillance is greater than ever as police forces are deploying facial recognition algorithms on massive scales with further threats to discrimination and injustice [143]. And these concerns are just the tip of the iceberg as IoT cameras have proven to be easily exploited [316] and computer vision techniques have developed that undermine privacy [249] and security [353].

Seeing these promises and concerns growing hand-in-hand, we must adopt techniques for comprehending and communicating these risks and steering technology away from worst-case scenarios. In short: we must figure out how to provide flourishing fields, such as computer vision, with a moral compass.

In this paper, we present a method we are calling *speculative analysis* to categorize, analyze, and communicate risks of emerging technologies whose final applications remain highly uncertain, such as those developing out of computer vision. Our approach brings together aspects of traditional risk analysis [328], speculative design/fiction [61, 28], and future studies [333] in order to garner foresight and take steps today that can lead technology down a more ethical path. We begin with an analysis of risk factors, categorizing current trends in CV by their potential harms imposed on society. This provides an overview risk characterization allowing us to identify vulnerable populations and the technologies implicated in risk exposure, as is done in traditional risk assessment [93].

Categorizing risk factors provides an organized vantage of our problem space, outlining what kinds of threats exist without a sense of the magnitude or likelihood. The goal of risk analysis is to characterize, communicate, and offer mitigation options for known risks [251]. However, traditional analysis methods have been critiqued for their inability to address uncertainty, consider risks from a societal lens, and take into account how non-experts relate to risks [174]. Since risks related to new CV technologies are not well understood, we aim to further our understanding by exploring scenarios. That is, much like designers do with fictional artifacts, we consider possible worlds that our technologies could create; each possible world being a projection of a risk framed as a technology causing a harm affecting a population. We then ask questions about these possibilities in a Bayesian manner: once the world has X technology, what's the likelihood of Y consequence. Rather than numeric probabilities, we can cast our probabilities into buckets of plausible, possible, and probable, justified by known factors about commercial desirability, technological feasibility, and ease of mitigation. Further, we compare our risks using a traditional 2D plot comparing factors of uncertainty with factors of severity as an assessment of overall threat posed [328].

With our risk factors categorized and scenarios compared, we can then focus on risks that are most likely and most threatening. To make further sense of these particular risks, we offer up narrative case studies that dig into the future contexts where these risks manifest. This task employs our creative capacities to think through what choices and changes must occur to allow a particular risk to develop. We limit ourselves to two narrative case studies related to CV risk meant to form communicative tools to discuss and analyze the risks we found to be most likely and most threatening. Using these narratives as tools, we discuss the assumptions made by these stories and what steps could be taken to avoid the described outcomes.

We conclude by discussing lessons learned, what further work could be done to improve our analysis, and how the field of CV may move forward utilizing such considerations.

4.4.3 Categorizing Risk Factors in Computer Vision

4.4.3.1 Bounding Our Considerations

Prior to characterizing and categorizing risk factors, a few framing assumptions and definitions must be made. Specifically, given the breadth of what may count as a risk analysis, the scope must be limited to a particular set of decision-makers and corresponding set of consequences relevant to them. What a business may consider risky is quite different than what researchers care about, which may also be askew from what the general public perceives. Thus, risk assessment is not meant to be a simple summary of science, but a method of enhancing practical understanding to guide decision-making of a particular group [93]. For the consideration of this paper, we will assume computer vision researchers and practitioners as our target group of applicable decision-makers. That is, we will analyze risk from a lens relevant to those deciding how to design, develop, and form best practices for new computer vision technologies. This means we will not, for example, consider the threat of damaging a brand image or harming business relations. Rather, we will focus on how decisions made by the community of researchers and engineers may relate to risks felt by the broader society, or in simpler terms, the users.

This brings us to further delimit our definition of risk. As risk researchers have established, the first part of defining a risk is to decide which consequences are included [137]. For the sake of this paper, we will define these consequences as harms or hazards that computer vision technologies may inflict on the public. Again, this is distinct from other definitions of risk, such as empirical risk embedded in certain methodological choices [260]. The assumption here is that the computer vision community cares to assess the societal impacts it may create and review practical considerations on how to avoid dangers and public detriment. Thus, we will limit ourselves to risks that enact a specifiable harm to an individual or group who might interact with a CV technology.

4.4.3.2 Categorizing Trends

Summarizing the broad progress in the booming field of computer vision has culminated in our finding five categories of risk that have CV-specific correlates and two general categories that any similar technology innovation may invoke. The five CV-specific categories are *privacy violations, discrimination, security breaches, spoofing and adversarial inputs, and psychological harms.* The two general categories are job loss and error/edge cases. Keeping our analysis specific to CV, we will overview each of the five specific risk categories, explaining what research and commercial trends create potential for harm.

Privacy Violations: This risk category is meant to cover all ways in which CV applications may lead to a third-party gaining unintentional or undisclosed private information about user. This may include, unwarranted surveillance, inferring information that was undisclosed, or de-anonymizing images. The potential for privacy issues appears pervasive given work currently emerging in CV research such as inferring health metrics from social media images [192] or de-anonymizing blurred images [249]. Further, due to continued progress in facial recognition abilities [53, 290, 185, 161, 49] the presence of any passive camera or image found online could easily lead to identification and potentially a privacy-violating inference. Technology ethics researchers Kate Crawford and Jason Schultz have termed the class of privacy violations that come from unanticipated inference as predictive privacy harms [96]. As our ability to make inferences from videos and images expands, so do the possibilities of diminishing trust and causing predictive privacy harms by inferring unintended and, potentially, consequential private information such as health informatics, uniquely identifying someone who did not choose to post a photo or video, or pinpointing a person's location.

Discrimination: Codified by our laws in place via Title VII of the Civil Rights Act, Title IV of the Higher Education Act, and the ADA, discrimination harms occur when someone receives unfair treatment due to their identity such as race, gender, class, or sexual orientation. Much like humans, the capacity to discriminate is further alive in machines. An undeniable trend in CV is

the heightened use of machine learning models, specifically convolutional neural networks [196]. The promise of machine learning is paralleled by the difficulty in making sure the resulting models are fair. Broadly speaking, the fear of bias and discrimination within AI and machine learning has become a topic of the day. Within CV, MIT researcher Joy Buolamwini has found that biased training samples have led to facial recognition models that do not work on black or other minority faces [73]. Further, research on age, race, and gender image classification continues to progress [203]. There is even work attempting to replicate models that can identify female attractiveness from a male viewer [350]. With racial and gender opportunity gaps being a continued problem of our time, technologists must not ignore how they may objectify or exacerbate these issues. Especially concerning are reports that racial bias is already showing up in mass facial recognition software used by police officers [179, 120].

Security Breaches: A large umbrella of risks imposed by CV technologies are varieties of security vulnerabilities. That are ways in which the presence or misuse of cameras or CV systems allow access to guarded information or systems. We classify a broad spectrum of attacks under this category. Propelled by CV innovations, the ubiquity of camera systems has opened up vulnerabilities any time proper security precautions have not been taken. Recently, massive attacks against CCTV cameras in Washington DC allowed up to 70% of security cameras in the region to be compromised [341]. Separately, ubiquitous cameras took a part in a large-scale IoT attack against DNS servers as botnets compromised hundreds of thousands of devices to be used in a one-off DDoS [316]. Further, researchers have shown that cameras can be used to steal information, such as passwords, off of filmed screens [353]. Without appropriate security, mobile cameras, home monitoring systems, web cams, and even calibration systems for critical systems have the potential to be co-opted by adversaries.

Spoofing and Adversarial Input: Broadly defined, this category, in the scope of CV, are adversarial attacks that attempt to get automated systems to react confidently to inputs while generating incorrect results. CV systems that are used for fraud detection, liveness detection, act as a security barrier, or have social consequences may be threatened by an adversary who

understands the system and can game it. There has already been a thread of research showing the ability to exploit deep neural nets [240, 330], soliciting high-confidence predictions for humanly-unrecognizable images. Other research has proven that these adversarial inputs are not simply laboratory scenarios, but expose a reality of real-world vulnerabilities [199, 156]. Researchers have already proven that this problem extends beyond well-understood systems. By targeting common CV tasks, such as segmentation and detection, it is possible to create systems that can generate adversarial examples against arbitrary blackbox CV systems [351].

Psychological Harms: Unlike other harms resulting from a CV technology's function, psychological harms are related to the wider effects created by ubiquitous cameras and passive monitoring. Having a world where personal devices, CCTV, drones, satellite images, and social media imagery are omnipresent and potentially smart (ie, actively making inferences) gives the impression of unending surveillance. This may lead to a constant state of stress and anxiety, and perhaps lead people to make social choices, such as not attending a protest, based solely on the fear of scrutiny or exposure. There is already a history of research showing that workplace monitoring leads to employees feeling more stressed [149]. Further work has shown that surveillance is likely to diminish people criticizing the government [85] or their ability and willingness to escape oppression [305].

Each of these five categories of harm should be of particular concern to CV research given their clear engagement with trends in the field and empirical understanding. While we do not spend time on them due to their generality to technology, job loss and unpredictable error are also aspects of CV research that should be taken seriously in their ability to negatively impact or harm society.

4.4.4 Scenario Evaluation

Relating trends in CV to categories of harm aids us in having a macro-understanding of risks; however, it does not give insights into which trends may be most worrying and where we should pay special attention. In order to get at these practical conclusions, we first must unpack these broad

trends into smaller components for analysis. Our unit of analysis is the scenario, which consists of a specific technological arrangement causing a harm to a population. The goal here is to entertain a wide range of scenarios to tease out the ones most worthy of deeper consideration. Utilizing scenarios for rapid, low-cost evaluation of technology has a deep history within design [28], future studies [267], and HCI research [248]. They are commonly used to test boundaries around norms [248], engage users in design processes [92], and enable analysis of future conditions that would otherwise be encountered with high uncertainty [333].

Our motivation was to generate scenarios that portray concrete future consequences that seem feasible given the identified categories of harm and application trajectories in CV. To generate the list, we asked ourselves and colleagues with domain knowledge to field examples of a way CV technologies might harm someone now or in the future. Some of the scenarios involve conjecture regarding future commercial applications and social states of affair such as image processing to determine job candidacy or to detect environmental hazards. For brevity, we will only discuss a few of these scenarios that are critical to our later analysis, but consider the full list in Table 4.1.

Throughout Table 4.1, you will find several different examples of how discrimination may leak into trained models, particularly CNNs. One scenario we discuss is that a training dataset containing an undiscovered bias gets passed around and used for varying commercial applications to only later discover bias. Given the uninterpretability of deep neural nets, if found too late, it could be impossible to fully remove the bias from the trained model. An escalated scenario relates the trend of police using facial recognition to identify suspected criminals. As is being discussed in current events [319, 143], facial samples used by police disproportionately sample African Americans due to historical bias in policing and crime. This could lead to a predictive policing system that uses threat scoring or, even further out, perhaps an autonomous security drone that monitors the public for criminals and has deep biases to suspect African Americans are criminals. This could objectify and exacerbate the very problem of prejudice we seek to eradicate in our society.

On a separate thread, we consider different scenarios where security breaches in IoT camera systems magnify. If a large enough botnet was successful, we could see internet outages that could

harm online and business infrastructure for large amounts of time. An even bigger concern would be if distributed IoT networks became a method to pass along more harmful viruses, much like the Stuxnet worm [361], searching for access points into vulnerable infrastructure such as the power grid or broadband systems.

Postulating ways spoofing attacks could turn awry, we consider scenarios where driverless cars could be attacked to trigger highway collisions by placing a carefully selected object in the visual range of the driverless vehicle. Separately, we could imagine CV being used to discover environmental hazards such as oil spills. If a malicious company wanted to cover up the event, they may tamper with the visual features of the hazard, say spraying a color-changing chemical onto the surface of the spill, to avoid detection by a known CV system.

Another scenario includes a privacy issue that may come with CV indiscriminately processing mass online photography. One way this could go wrong is if photos are posted online without someone's consent that then get processed, tagged, and finally associated with a profile that costs them a job or harms their personal life. Similarly, we could imagine CV used by insurance companies to better assess the health of applicants. This may allow a photo that was never considered relevant to a health care to cause a spike in someone's insurance premium or even show evidence of a pre-existing condition that was unknown or untreated.

4.4.4.1 Likelihood Analysis

While these scenarios range from sci-fi imaginings to already-known problems, we believe they all represent issues that are within the realm of possibilities for a future state of affairs. Through a Bayesian lens, $P(Scenario|Technology)$, we would not weigh each equally. Taking into account the commercial pressures to see CNNs revolutionize our ability to handle Big Image Data along with the quick spread of ubiquitous camera systems, we believe discrimination, large-scale security breaches, and damage to democracy through psychological harm to be among the most probable concerns. To reach this conclusion, we took each scenario and categorized it into either possible (meaning it's technically feasible, but easily avoidable), plausible (meaning it's feasible, hard to

avoid, but would take a very malicious actor), and probable (meaning it's feasible, hard to avoid, and already on the path to occurring). Using an adapted visual aid developed by designers Anthony Dunne and Fiona Raby [28], we show how each of these scenarios rank in Figure 4.1.

4.4.4.2 Plotting Risk Factors

The next part of our process involved plotting these scenarios along dimensions used in traditional risk analysis to compare factors of **uncertainty** and **severity**. We borrowed our metrics from a canonical risk perception study [328] published in 2005 to rank risks as perceived by different groups of the public. In the original study, they included risks ranging from motor vehicle accidents to nuclear power to vaccinations in a major study done over years to assess how populations compared different societal hazards. For our analysis, we are only considering risks within CV technologies to keep a tight domain focus.

To create a metric for **uncertainty** we assessed the following factors: observability (is the harmful effect easily observable?); newness (is this a new risk or one society has long faced?); known exposure (does a person know they were exposed to the risk?); scientific knowledge (is the risk well understood by scientists?). The less known and observable, the newer and more difficult to infer exposure, the more positive the value for uncertainty. The metric of **severity** was structured by a separate set of factors: controllable (do practitioners have a lot of control over the risk?); detrimental (are the harms common or detrimental to the population involved?); scale (do the harms occur at a large, global or small, individual scale?); risk to future generations (are the effects lasting burdens on future generations or quickly addressable?); mitigation (is the risk easily mitigated or difficult?). The less controllable, more detrimental, larger scale, more of a future burden, and harder to mitigate all contributes to a more positive value.

We gave values to this 2D metric based on facts about harms we know through the news, mitigation tactics published by field experts, and how much of a damage the final harm wages (ie., embarrassment $<$ financial loss $<$ physical harm $<$ detriment to societal functions). While we believe this is a fair way to embark upon an initial overview risk assessment of the field, thorough

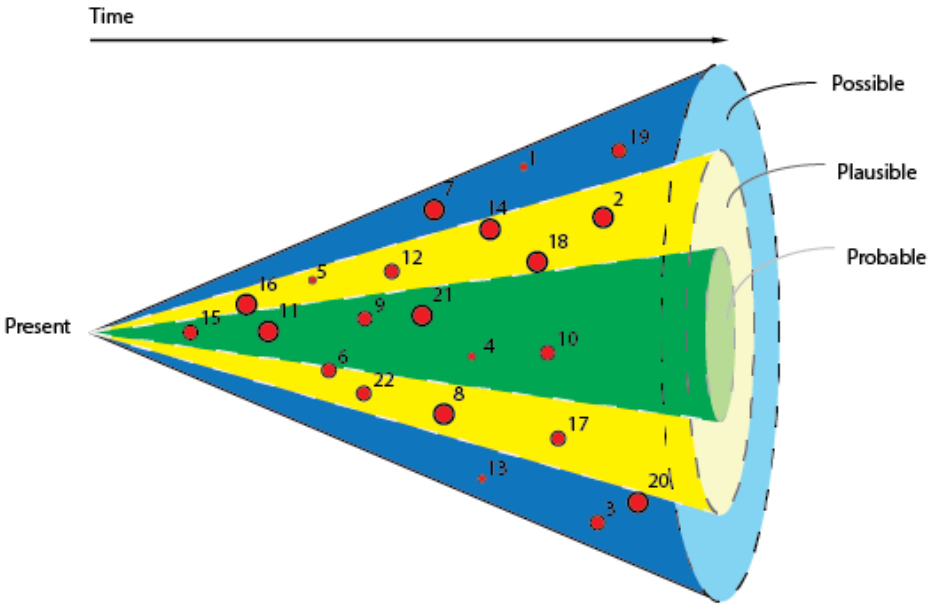


Figure 4.1: Likelihood categories of possible futures.

follow-up studies could be done allowing expert and non-expert populations to weigh in on how they rank and compare risks along these same metrics. Using the above questions and assessment criterion, we constructed the plot appearing in Figure 4.2.

We will elaborate on how we arrived at some of our highest ranking risks (ie, top right quadrant of Figure 2). Scenario 2 - people afraid of protesting because of surveillance and face recognition - was treated as a vast concern. Not only is it very difficult to observe the actual psychological distress that could cause this, it may occur over a longer period of time as various punishments accumulate. Also, it could vary drastically depending on who is in power and how laws progress around information sharing. As Frank Pasquale discusses in The Black Box Society, Occupy Wall Street already suffered from some of these conditions as Wall St. banks, unknown to activists, gave security camera footage to the FBI, allowing certain protesters to be identified and targeted [254], likely with the aid of face recognition software. Further, once this change occurred it would impact the ability of future generations to change the status quo making corrupt regimes even more powerful.

Scenario 10 - police unjustly searching and arresting people of color due to a bias in visual analytics done on surveillance and camera monitoring systems - was ranked as both severe and uncertain. Uncertain because a single instance of a person targeted as a threat by a visual monitoring system would not necessarily augur this deeper issue. It may take years of injustice and expert assessment of the systems to fully comprehend the risk. In the meantime, the consequences to human lives would be severe and it could breed further distrust between communities, destabilizing social foundations.

Scenario 21 - IoT cameras carrying along a virus hoping to pass it to vulnerable infrastructure - was assessed as a severe concern. Given the difficulty in controlling security vulnerabilities multiplied by the instances of devices that continue to be networked online creates an exorbitant concern. On the other hand, we did not rank the uncertainty so high. While the attack vectors may grow in size and it's hard to tell if a system is infected, the idea that networked infrastructure needs strict regulation is not new. The use of networked devices to monitor and control infrastructure

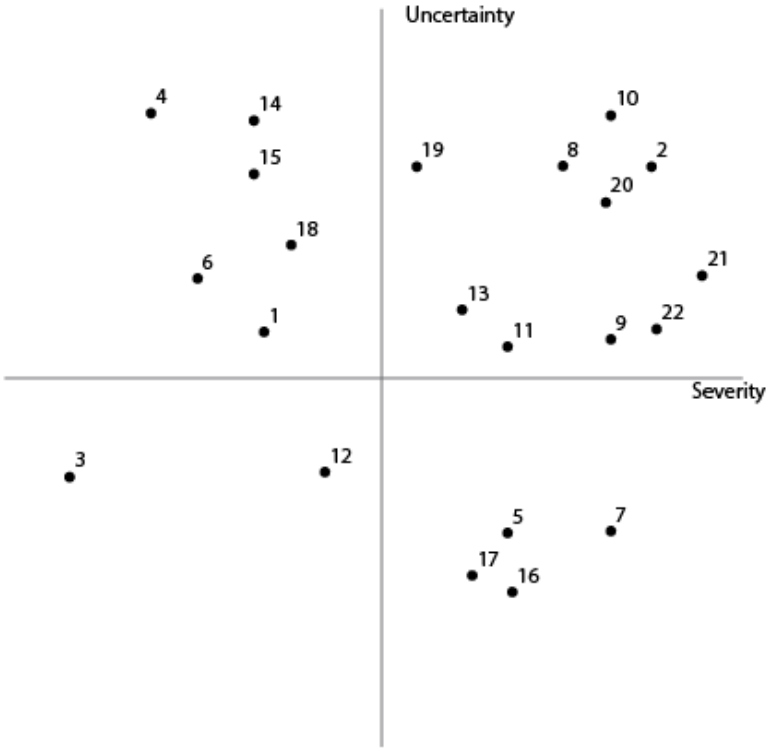


Figure 4.2: Uncertainty vs. Severity for 22 CV Risk Scenarios

has increased over decades allowing cybersecurity experts plenty of time to consider attacks. With that said, new adversaries and vulnerabilities continue to emerge and the scale of the security concern should raise a red flag. A well-planned attack in this realm could involve multiple CV-based attacks, first passing a virus, then exploiting a CV verification system, or even spoofing a biometric monitor for a security officer.

4.4.5 Narrative Case Studies

Narrative and fiction are commonly used constructs for research and exploratory purposes with HCI [61, 58, 307] and design theory [28]. The added value of narratives over simply adumbrating scenarios or analyzing fail modes of systems is that they create a context that is better able to represent social or political conflicts [60]. That is, locating the consequences and frictions of engineering decisions is often difficult in purely technical descriptions of systems where accuracy and efficiency, system dynamics, and usability are often analytic constructs that lend to objective solutions. However, when we consider harm to people, we need richer depictions that allow us to consider thorny matters such as social norms, notions of justice and fairness, and trust. As DiSalvo argues, fictional examples offer opportunities for interrogation and challenge [112]. A good technological narrative should structure a place for discussion, disagreement, and ethical deliberation among experts. While we do not have the space to construct narratives for every scenario nor elaborate on details of future worlds, we offer up two provocative flash fictions (Figures 4.3 and 4.4) to deconstruct, analyze, and add to the repertoire of conversation among CV researchers

In the Scenario 10 narrative (Figure 4.3), we see two men discussing an event where one was arrested due to a police confrontation instigated by a camera system that identified him as a threat. The implication is that society has committed itself further to the utility of smart camera systems, trusting the inferences they make to guide police response and create efficiencies in physical security. Presumably connected to a vast database of faces and operating with an AI model that can cast likelihoods of someone's intentions, the inferences made by visual data have allowed discrimination to move into a more objective realm. As stated by the protagonist, he was targeted for reasons

Ray exited the correctional facility a free man. The long walk between the brick-and-mortar structure and the gated entry was lined by swiveling cameras. Sentries were replaced by intelligent observation systems - Watchful Eyes, as the policing community called them. Being the same system that got Ray into trouble six months ago, he and his brother couldn't shake a looming sense of discomfort as they eagerly hopped in a car to escape the computational gaze.
"Man, it's good to be out of there. They call this a free society when you wind up behind bars for six month for nothing. Walking to pick up my kid at school and next thing you know. How's my daughter doing? Thanks for watching after her."
"Traumatized no doubt. She don't want to be around computers anymore I can tell you that much. Sit a phone on the table and she'll put a napkin over it to cover the cameras. Tells people "computers" took her Dad."
"Humans took her Dad. Computers told them I was a danger. That's why it took my case six months to get dismissed. The officers were "working with the information they had" is what I was told."
"You think it would've been different if you stayed calm?"
"I don't know man. You're telling me I have to give in, let the computers oppress us now? I'm three blocks from my daughter's elementary school and I get cut off by a cop car. They ask for my ID and tell me they're gonna pat me down. All because one of those cameras saw my face and decided I was a threat? You better believe I'm losing my temper in that situation."
"So those cameras are broken, is what's up? That's why they let you out? What's that even mean?"
"Yeah man. Algorithmic bias, they call it. Apparently no one warned these programmers that past examples might teach their little computers to see a black face and think criminal."
"Well stick with this class action suit your lawyer got you in on. Take a small chunk of those giant paychecks these technology companies receive."
Ray and his brother went silent as a cop car with a mounted camera crept past them on the highway.

Figure 4.3: Deeply Learned Bias (Scenario 10)

"Take all home surveillance systems offline," they were told. Until security experts could come up with a solution, people were forced to power down their networks of devices. Up until last week, the dispersion of networked devices around the home were a boon. Your coffee was ready when you woke up, your thermostat adjusted itself to the erratic weather patterns, and faithful cameras watched over your home, products, and in some case children. Sarah sat in her shadowy home, waiting for the phone to ring for an interview with the New York Times about her experience in the blackout. Power was back up, but she preferred to leave as much equipment off as possible since she didn't really understand which devices could be hacked or not.
"Hi, Sarah? This is Preet Singh with the New York Times. Is now a good time?"
"As good a time as any."
"OK, I'm going to start recording now, please let me know if there's anything you'd like excluded from print. Tell me a bit about what happened when your power first went out."
"Well I was arriving home from work and normally my garage just opens for me when it sees my car or license plate or however it works. But it wouldn't open. So I went on my phone and tried to open it manually and that did nothing either. Only then did it hit me that it must have been a bigger outage since the red light down the block was out too. Being a bright and sunny day, I was very confused and to be honest scared."
"That's understandable. And what can you tell me about what you've learned since then about the situation?"
"Well from what I can tell, I got the least of it out here in the suburbs. It was mayhem in the cities. Now I don't really understand the details, but apparently my home security system, thermostat, everything really, might have participated in the attack. Something happened here for sure. Many fuses in my home were blown out. I guess moments before I arrived home everything surged. Is that right?"
"That's right. What's known is there was a large-scale exploit that started with camera systems connected to the internet. But now that these cameras are connected into integrated homes such as smart thermostats and lighting systems, they were able to create timed surges that targeted certain distribution assets in the power grid."
"Oh my.. So someone used a camera to operate my home?"
"Essentially, yes."
"We're always trying to advance so fast, told to buy the next product. Doesn't anyone test these things first?"
"Of course ma'am, security threats are known, but are very hard to control. No one saw this coming, I can assure you."
"Well I believe that, but city-wide blackouts, my goodness. Whoever thought of this stuff should've warned business before they released so many products."
"Do you believe an event like this will change your future trust in technology products?"
"Absolutely. How could it not? Is all this really worth some minor convenience?"

Figure 4.4: The Cameras Attack (Scenario 21)

that were discovered, much later, to be related to a racial bias within the system. We implicitly understand that he must have been an innocent person meaning it is unlikely the system was connected to any human-in-the-loop overseer who may have been able to redirect the police who responded. Further, he informally suggests that the history of racial bias in policing should have made programmers anticipate severe biases in any system trained by that data.

How far off is this kind of scenario? And how dangerous should we see it to our society? The idea of predictive policing is already on the rise. Data-driven approaches and machine learning applications are currently being tooled to predict crime [225, 179, 2]. Guessing the likelihood of recidivism [43] and setting bond [27] are further becoming areas where computer science is at work, and already signs of racial bias are showing up [120]. With computer vision research emerging that attempts to predict criminality of a face [349], and the many advances of object detection and face recognition, it seems quite likely CV's application to policing will continue to grow. Of course, the goal of the research community should be to diminish bias rather than exacerbate and obfuscate it. One must also understand that, while a prejudice police officer adds harm and reduces trust, this person can be isolated and ideally, punished. If a widespread vision system was found to be biased, the implications to trust could easily fall on an entire industry with the camera acting as a symbol. In an industry that already has systematic disparities in demographic representation [243], care should be taken to ensure that the applications, training sets, and best practices avoid the growth of such a scenario.

Scenario 21 (Figure 4.4) presents us with a working woman who was impacted by a major attack against the power grid. Seeded by vulnerable security cameras, a worm made it to seemingly millions of homes to create timed power surges. This was made possible due to a supposed advancement in integrated or smart homes. The story suggests that many homes have become equipped with intelligent camera systems, thermostats, and appliances, allowing an attack on any one system to be threatening to the whole ecosystem. One assumption is that home network security has not improved significantly in the interval between now and the context of the fictional tale. It is also taken for granted that people continue to be lax about data sharing between devices and

systems. Given that a camera system could act as a critical part of my other smart-home devices, we assume the camera would be connected to nearly everything in the home. We also conjecture that an adversary who understands the intricacies of the power grid could also design an adversarial system that could precisely surge power in homes, placing a critical load on particular assets.

How far out is such a scenario? To what extent does it really implicate CV researchers? It should immediately strike readers that leveraging insecure, distributed devices is a reality of our time that is unlikely to go away. On Friday October 21, 2016, we saw the largest DDoS attack ever, using IoT devices, particularly cameras, to deliver 1.2Tbps targeted at DNS provider Dyn [285]. Security experts have warned that these attacks are likely to grow in size and frequency [10], and that the market is not the place to look for solutions [285]. What makes this issue particularly tricky for CV practitioners is that unless the computer power required to perform video processing significantly diminishes, most systems using a video feed will require internet access. That is, much like Google's NLP engine relies on cloud services to process audio samples, CV seems destined for a similar future in the cloud. In the long-run, large-scale attacks could cause both blanket harm to society and mass distrust for using cloud-enabled or IoT devices. A moment such as the one described in this narrative would likely signify the necessity of government intervention on the problem which, if impulsive, could severely deter industry development and limit the applicability of research insights.

4.4.6 Conclusion & Future Work

Emerging from this dive into the risky situations that CV research might lead us into, we see a number of takeaways applicable to further develop ethics in this field. To begin with, we have postulated and explored a variety of scenarios where some sort of disparate, yet significant impact is enacted through bias and discrimination. Preventing this reality will take a lot of work, but has tangible ways forward. One way forward could be to seek professional certifications for particular practitioners who design systems where the results have serious life consequences. Much like a certifications for doctors, lawyers, architects, and professional engineers, we could see sub-fields of

computer science adopt licensure programs. Something that can be done sooner is taking seriously our responsibility to perform blackbox tests, audit our systems, and provide access to unbiased datasets. These efforts already have early starting points with the Algorithmic Justice League and the Fair, Accountable, and Transparent Machine Learning Conference.

Another place we may consider diverting expertise into is both for- and non-profit oversight projects. Much like cybersecurity has pen-testers who work toward bug bounties set out to prevent major hacks, we could imagine adversary bounties where researchers prove the ability to create adversarial examples to systems before they go live. Similarly, we could see public-interest groups who certify particular systems as fair, using their seal of approval to aid the public in choosing systems developed by best practices.

Other mechanisms that may help mitigate some of these risks are working sooner, rather than later, with policy experts to advocate for security standards on IoT devices and routers. The more distance between experts and policy-makers, the higher likelihood the eventual policies designed will be damaging to the field. In the same vein, to prevent some of the concerns around privacy and de-anonymization, there are already regulatory models, like the EU's GDRPs, that attempt to give users more control and knowledge over who owns and uses their data. While this could be seen as a short-term inefficiency to data mining operations, it may prevent a long-term turn away from the field as the abundance and severity of privacy harms develop. Last, but certainly not least, is experts weighing in on the kinds of systems that should keep a human in the loop. It is exciting to see the accuracy and capability of CV work grow, but it is critical that practitioners recognize the limits of what sorts of judgements we want automated and where checks and balances should exist.

As a broad effort, this work points easily toward further deep research on the topic of risk in CV. Surveying more experts in the field about risky scenarios and future applications could enrich the assessment and help the public and policymakers understand what emerging trends are most dangerous. Further, given the interrelationship of psychological harm, trust, and technical knowledge, information could be gathered from users of these systems to get a more targeted

assessment of how different populations perceive these risks. Finally, CV researchers, and computer scientists at large, should actively determine how they can make ethics a central part of their concentration area. Incentivizing ethical innovation must be a major factor to any serious interest in warding off dangerous or harmful problems in an expert domain. Emphasizing the ethical dimensions of CV research and taking seriously the study of risk factors such as those discussed in this paper will ensure a prosperous and fair future of the field.

#	Scenario	Technology	Population	Harm
1	Health insurance premium is increased due to inferences from online photos	Biometric inference	Any individual	Privacy
2	People will not attend protests they agree with due to fear of recognition by cameras and subsequent punishment	Face Recognition	Public	Psychological
3	Person is unfairly denied entry into public location due to visual scan at door	Image classification	Minority community	Discrimination
4	Person denied job for photo they did not post online that was de-anonymized by CV.	Face recognition	Any individual	Privacy
5	Security guard sells footage of public official typing in a password to allow for a classified information leak	Key stroke or screen inference	Any individual	Security Break
6	Job candidate is denied job because classifier of "attractiveness" was used within a model.	Social psychology classification from image	Women	Discrimination
7	Environmental hazards tracked by aerial imagery are missed because company tampers with visual appearance of a pollutant	Machine learning	Public	Spoof/Fraud
8	Automated public transportation that uses visual verification systems is attacked causing a crash	Driverless cars	Public	Spoof/Fraud
9	Xenophobia leads police forces to track and target foreigners	Facial recognition and image classifiers	Minority populations	Discrimination
10	Police unjustly search and arrest people of color due to criminality inferences.	Human inference from images	Minority populations	Discrimination
11	Online infrastructure is brought down through IoT attack using hacked cameras	IoT cameras	Public	Security Breach
12	Programmer uses third-party CV model to create eye tracking tool that, at release, does not work for Asian faces	Eye tracking	Minority population	Discrimination
13	Autonomous security system using CV, incorrectly detects object as weapon, leading to unjust attack or arrest	Object detection	Anyone	Error
14	Corporate ethics spiral as whistleblowers are deterred by workplace monitoring	Camera surveillance	Public	Discrimination
15	Automatic captioning leads to thousands of offensive captions on public photos	Image captioning	Any individual	Discrimination
16	Anonymized health dataset used for CNN training gets de-anonymized by adversary, revealing health info of millions	CNN and de-anonymization	Public	Privacy
17	Recreational drones for extreme sports video popularizes and incidentally captures videos of children in public spaces	Drone cameras	Children	Privacy
18	Death of private life - people must assume all matters of life may be used against them in work, court, etc	Ubiquitous camera systems	Public	Psychological
19	Disaster response dictated by aerial imagery ends up sending all first responders to rich neighborhoods because classifier uses inferred value of property	Image classifiers	Low SES Populations	Discrimination
20	Automated weapon is triggered to attack innocent people due to adversarial attack against visual processing system	Thread modeling from images	Public	Spoof/Fraud
21	Mass IoT network is used to pass a virus along and deliver into public infrastructure system, taking down portion of power grid by creating timed surges	IoT Cameras	Public	Security Breach
22	A popular image dataset gets used to train dozens of commercial CNN applications, and is discovered to have a major bias in it that disadvantages minority groups	CNN	Minority Group	Discrimination

Table 4.1: Twenty-Two Risky Scenarios Used for Analysis

Chapter 5

Technology Ethics in the Public Sphere

5.1 What's at Stake: Characterizing Risk Perceptions of Emerging Technologies

5.1.1 Prologue

Motivated by reading research in the literature of risk perception, this section looks at survey research done on how different groups perceive the risks of emerging technologies. Primarily studying experts and non-experts, miniature risk scenarios were used to describe potential consequences of emerging technologies. Thus instead of simply asking participants about "bias algorithms," we used slightly more developed scenarios such as "a biased algorithm determining someone's qualification for a job." These were not fully narratives using characters and worlds, per se, but offered **templates** that were thick enough to describe specific harms. However, to enrich the results, we did ask participants to describe their best and worse-case scenarios of their top risks by ranking. The use of narrative here allowed us to gain clarity on the different mental models participants were using when interpreting our risk scenarios. This help validate our findings and compare different ways certain populations thought about the problems at hand.

5.1.2 Introduction

Emerging data technologies are primarily developed by capturing or acquiring a large data set to use for further analysis or model training. Designers also build responsive and personalized systems that learn from user behaviors. In both cases, these data are generated from user-specific

behaviors tracked and archived, often on public, web platforms such as Facebook, Twitter, etc. Thus, users largely lay the foundations for the accelerating area of Big Data Technologies, including machine learning (ML), artificial intelligence (AI), and behavior-driven design, and these users must rely on the companies and parties to whom they have given their data (knowingly or not) to be ethical.

Yet, we already know that many impacts (e.g., privacy, ethical, legal) and constraints (e.g., protocols, technological capabilities) of online technologies are poorly understood by users [223, 52, 323, 130]. We also know that, when asked, users are often uncomfortable or find undesirable the practices of online behavioral advertising (OBA) and personalization [324, 320]. This misalignment is often framed as a consumer trade-off between privacy and personal benefit [121, 356]. Framing it this way leads to an assumption that the benefit of web services must outweigh consumer's privacy concerns since users are not opting out of services.

However, if consumers really are performing this cost-benefit analysis and making a conscious decision, then why we do we see such hype and panic around risks and harms caused by technology in the media? Daily news headlines relay injustice [146, 5, 27, 312], personal boundary violations [301], and gloom [230, 145, 123] over the impacts of technology on society. Some of these problems may indeed warrant concern from the public and social advocates; others might be overblown headlines to keep technology followers clicking. Meanwhile, previous research shows that, when prompted, users do have concerns over issues like data privacy [55, 187, 35]. Therefore, we wonder, do users actually understand the relationship between the data they hand over and the systems being built from that data, including those that could lead to the kinds of risks and harms they read about in the media? What dimensions of how people perceive risk impacts their overall judgment? Are these issues being communicated equally across different pockets of the public?

In order to better understand the psychological dimensions of this problem, we modified and implemented a classic survey instrument from risk perception literature [296]. The original study validated an interesting phenomenon: that the more voluntary one perceives a risk to be, the more willing they are to accept it. The paper also examined the differences between expert and non-

expert risk perceptions. We designed a derivative instrument focusing on risks posed to society by emerging technologies (rather than the original study's focus on environmental and health hazards), with a particular emphasis on hazards from technologies emerging due to the rapid availability of data.

After surveying 175 people, 26 of which were technical "experts," we found significant differences in their perception of risks related to technology. Our findings indicated interesting contrasts between how the groups ranked and rated risks. We also found that many risks related to data technologies were perceived as involuntary by both groups.

These results have implications for design, public communications about technology, policy, and the study of user mental models. In this paper, we discuss the relevance of studying risk perception to HCI, report findings from the deployment of our derivative instrument, and analyze these findings in light of a simple model we are calling "risk sensitive design" for interpreting when misaligned risk perceptions may warrant reconsideration.

5.1.3 Related Work

5.1.3.1 Past Work on User Attitudes and Beliefs

A lot of complexity lies in the details of how data is processed and shared once a user opts into a service that captures behavior and information. And while there are benefits for personalized features, numerous studies have shown that users do not always want or like these perceived benefits, or that they do not think these benefits outweigh privacy or related concerns [29, 320]. For example, users think that data being sold to third parties does not benefit them [55].

While some may argue that this is simply a logical trade-off of privacy for convenience or utility [356], research has shown that feelings, comprehension, and reasoning patterns users apply in real situations do not actually support this baseline privacy trade-off. Studies by Cranor et al. have shown that users carry nuanced views about the context in which they feel comfortable with information disclosure, and that they may dislike true data practices, if properly understood

[212]. Besmer and Lipford have furthered findings that once misconceptions are clarified, users may regret their technology and disclosure choices [52]. Other studies have shown that it often takes a personal experience before people elevate their awareness of risks they may be ignoring with online technologies [187]. Problems like this could be related to design choices. For example, when given graphics on their mobile devices about when data is collected and where it goes, users are often shocked [35]. Findings by Angulo and Ortlieb point toward the need for better design to educate and inform users around the scenarios they find most concerning [26].

The issue is, we do not always know where these nuanced anxieties truly come from, particularly since, rather than understanding how technologies work, individual users are often applying their own folk models to reason about technology [339, 358]. This makes it difficult to discern ideal forms of transparency in design. We do, however, know that this transparency is not being achieved by the legal agreements that should clarify risk. A number of researchers have shown, for example, that privacy policies are too complex for the average adult [213, 175], and that users often misunderstand what rights they give over to platforms in their content under the copyright policies to which they are bound [135]. Barocas and Nissenbaum further explain [42], that given the complex nature of our data economy, it would be nearly impossible for these fixed agreements to be truly transparent. The legal and technical sophistication of terms of service and data markets make the idea of solving the problem with a unilateral, transparent opt-in agreement a hard goal to chase. However, given the heightened importance of technology and data to all of our lives [321], it is imperative we continue to innovate on how to make users more aware and shape the future of technology to equally benefit everyone.

For these reasons, we believe there is interesting work to be done looking not at what users comprehend from a terms of service, privacy policy, or a particular interface, but instead what risks they perceive related to technology broadly. As research by Acquisti and Gross suggest, privacy concerns may not correspond well to user behavior [15]. Thus, it may be that studying other factors such as risk could clarify dimensions of user reasoning and behavior patterns.

5.1.3.2 Risk Perception Studies

There have been many interesting studies within the risk perception literature that could be relevant for HCI and technology policy research. The early goals of the field resonate with some the current work studying user attitudes and believes. A foundational paper, "Characterizing Perceived Risk," states the problem as follows: "[This risk perception study] aims to discover what people mean when they say that something is risky; to develop a psychological taxonomy of risk that can be used to understand people's perceptions and predict societal response; and to develop methods for assessing public opinion about risk in a way that is useful for informing policy decisions" [297].

This literature has often tried to characterize how a balance is reached between perceived risks and perceived benefits [294]. We believe this framing may help designers and policymakers working with socio-technical systems. Particularly relevant to HCI, risk perception research shows a higher willingness for people to take on risks when they believe they are voluntary [296]. That is, risks such as driving or football are taken with less concern because the individual has chosen to do so. They also found that the differences in risk perception between experts and the public lead to concerning trends in inter-group trust and democracy [295].

Seeing these past findings as having relevance to today's conversations about user perceptions of computing systems, we believed there is value in replicating a risk perception instrument that focuses on technologies germane to contemporary society. Next we discuss the shape our adapted instrument took and our methods for seeking and grouping participants.

5.1.4 methods

Our interest in applying work from the risk literature to HCI was to garner fresh insights about how users and experts think about the risks posed by new technologies. The original study we replicated came from a 1980 paper called "Facts and Fears: Understanding Perceived Risk" [296]. In the write-up they summarize several studies that worked with a particular instrument for comparatively ranking, individually scoring, and psychological characterizing a long list of risks.

With over 1000 citations and a number of validations of prior research in that literature, we were interested to apply it to technologies that are now of high relevance to computing researchers.

The first iteration of their instrument contained 30 risks that participants comparatively ranked from most to least risky, and then also provided a raw risk score on a scale from 1-10 (from "very risky" to "not risky"). Participants made estimates on annual fatalities from the different risks and characterized each risk using 9 psychological factors. A later version of this instrument deepened this analysis by using 90 risks and 18 psychological factors.

Risks included were a mixture of common and technical risks–e.g., smoking, handguns, x-rays, contraceptives, vaccinations, and pesticides. The psychological factors, drawn from prior risk perceptions studies, included voluntariness of the risk, immediacy of effect, knowledge about the risk both individual and by expert scientists, control over the risk, severity of the consequence, the commonality of the risk, and familiarity with the risk. The authors surveyed a group of 15 "experts" who worked in relevant fields such as medicine, risk analysis, and policy, and compared their perceptions to other groups–a sample of students, a sample from the League of Women Voters, and sample from an active outdoors club.

Our adaptation of this survey truncated the size of the instrument–using 18 risks and 6 psychological factors–to keep the survey time to around 20 minutes and reduce the cognitive load required of our participants. We kept in 3 of the original risks for benchmarking and comparison,and then adapted the other 15 risks based on emerging data-driven technologies that have been discussed extensively in research and current events. Starting with a list of 30 risks, we shared our initial choices with colleagues working in large tech companies and research institutes, to make sure we chose the most relevant and interesting risks. In the end, the 15 new risks we used were highly relevant to today's reporting on technology issues. "Death or destruction from autonomous drones," "biased algorithms for filtering job candidates," "identity theft", "security breach of an online account," "filter bubbles (individuals receiving different versions of the internet)", "technology divide (technology only benefiting a small elite)", "job loss from automation," "nude photos being leaked," "having your online activities researched without consent," "Distributed Denial of Service

Attacks," "Undisclosed third-parties having access to your data," "Online bullying and harrassment," discriminatory algorithms used for policing," "hacktivists leaking large data sets containing personal information," and "malfunctions from driverless cars" were the added risks. From the original survey we kept "nuclear reactor meltdown," "harm to one's health from vaccines," and "plane crashes" which were picked from the top (nuclear reactors), middle (plane crashes), and bottom (vaccines) of average rankings from the original study.

Similar to the original study, participants ranked the 18 risks comparatively and individually. They also characterized the 15 new risks using 6 psychological factors along a 1-7 scale. For our study we used, "voluntariness," "fear of risk," "severity of consequence," "perceived self understanding of risk," "perceived understanding of risk by domain experts," and "likelihood of risk happening," where a 1 minimized the factor (e.g., "involuntary," "do not fear risk at all") and a 7 maximized the factor (e.g., "completely voluntary," "a deep fear of the risk"). In keeping with the goals of the original study, we framed our survey as involving "risks posed to society," and explicitly asked our participants to rank these risks as a threat to people broadly not merely to themselves.

In order to replicate the initial survey's distinction between "experts" and "non-experts," we distributed the survey to multiple populations and made this distinction based on career. For the current study, "experts" constitute people with careers or focuses in computing and technology development. We piloted the survey with both experts and non-experts from our social networks, and iterated on the description of the risks, as well as the wording of survey questions, based on their feedback. We also used pilot responses to tweak the timing of the survey, to ensure that it would take less than half an hour to complete.

Targeting experts, we deployed the survey to people in tech careers via snowball sampling following initial seeding within the first author's social network. To target a general population of non-experts, we also deployed the survey using Amazon Mechanical Turk (mturk), a crowdwork system where workers complete small tasks for payment. Though there are known biases in the mturk population that prevents broad generalizability, it has been shown to have advantages over

localized populations for demographic diversity [227]. However, the population does tend to skew younger and less ethnically diverse [167]. Though prior work has shown that mturk workers tend to be slightly more tech-savvy than the general population [134], they also skew less educated overall than most American working adults [167]. In the next section, we report the demographics from both expert and non-expert populations, so that our results can be interpreted with these factors in mind.

We deployed the survey on Mturk in June 2017. Based on our pilot testing, we were aware that the survey took between 15 and 30 minutes to complete. We paid workers $3.00 for the task, ensuring that the rate of pay would typically be over minimum wage. We included two attention checks in the survey (asking for a certain answer, to ensure the questions are being read), a strategy which is used to help ensure the validity of survey responses on mturk [103]. 166 workers completed the survey, and we removed 17 responses for failing attention checks or providing incomplete data, resulting in a total of 175 responses. As part of the survey we also had data about participants' careers. For those who fit our definition of "expert," we removed them from this dataset and added them to the expert dataset. The final result is 26 expert responses and 149 non-expert responses.

In addition to quantitative measures that were drawn from the original risk perception survey instrument we also added two open-response questions to the survey. First, for the three items that they ranked as the riskiest (i.e., the top three), we asked them to describe what they thought the worst case scenario was. Additionally, we asked if there were any additional serious risks to society caused by technology that we had not asked them about. During our analysis stage, we analyzed these responses qualitatively, using iterative, open coding [283].

5.2 Results

Analyzing survey responses, our goal was to better understand how people perceive different risks. That is, we view risk as an operational construct within human judgment to reason about certain decisions, such as when to act or not and when a situation is threatening or not. People can therefore only weight these decisions based on what risks they perceive. Our results section

starts with an overview our population sample, then highlights group differences for comparative risk ranking, raw risk scoring, and psychological factors as they correlate to perceived riskiness.

5.2.1 Properties of Population Sample

Our sample consisted of 175 completed surveys after removal of participants who (1) did not pass our attention checks, and/or (2) did not complete the survey. Following findings from the original study, we primarily analyzed the difference between expert and non-expert populations. Expert was defined as someone with either a career in computing with a primary role that is technical in nature or a student seeking a degree in a technology-focused field such as computer science or electrical engineering. Using this distinction, we separated our sample into 26 experts and 149 non-experts. The goal of this grouping is to develop an understanding of whether experts perceive technological risks differently than non-experts, given the often complex and future-oriented nature of these risks. Demographically, the sample included 96 men, 77 women, and 2 non-gender identifying individuals. Along ethnic lines, we split our population into 124 subjects who primarily identified as "White" and 51 subjects who primarily identified as a non-white ethnicity. Within groups, demographics were less spread. Our expert group of 26 had 25 men and 1 non-gender-identifying individual. 20 out of the 26 were white with 2 black, 1 hispanic, and 3 mixed-heritage respondents. Our non-expert group contained a wider spread with a nearly 50/50 split between men and women (82 women and 80 men). As expected, our differed widely on educational background. The expert group contained 30% with a masters degree, 46% with a bachelor's degree, 11% with a doctoral degree, and the remaining 3 reported some college (2) or a high school degree (1) but still claiming to work in a technical role. Our non-expert group contained 5% with a masters or professional degree, 44% who had a bachelor's degree, 9% an associates degree, and the remaining 42% having a high school diploma or some college, but no degree.

Though we sought to replicate a prior study that had a small expert group (N=15) to compare against, we were concerned about the cohesion of our expert group given its size (N=26). Though this could be even further improved in future work, we did see somewhat tighter agreement among

experts than non-experts. Within items asking respondents to compare risks, non-experts had an average standard deviation of 4.79, while experts had 4.16. On items that asked for a raw 1-10 ranking of risks non-experts had an average standard deviation of 2.76 compared to experts' 2.39. We further noticed a general trend that those in the expert population tended to rank technology risks much higher compared to common risks (e.g., plane crashes, nuclear reactor failure) both comparatively and using a raw score, whereas non-experts were consistently more concerned with these common risks. We will characterize this finding further below, but we took these as signs that there was a coherent enough difference between populations to garner insights.

5.2.2 Risk Ranking and Scoring

As Figure 1 indicates, the comparative ordering of risks (from most to least) was quite different between experts and non-experts. The top of the list looks similar, though non-experts were more concerned with identity theft (p=.063; two-tail t-test) and experts with job loss (p=.068). Average expert rankings showed the top three risks to be (1) job loss, (2) account breach, and (3) identity theft. Non-experts had the same top three items, but in different order: (1) identity theft, (2) account breach, and (3) job loss. This makes sense, given that experts could believe themselves to have a more sophisticated ability to control the identity theft, but not mass job loss. However, it is also notable that these three risks are well documented in media and have a national attention surrounding them.

With respect to participants' 1 to 10 risk score for each item (see Figure 2), the results generally validated comparative rankings; however, we did gain more nuance in terms of the difference in magnitude between certain risks that would have been unclear from rankings alone. Figure 2 plots risk scores comparatively with average expert scores on the x-axis and average non-expert scores on the y-axis. This data shows, in general, non-experts perceive all risks as greater in an absolute sense. With 1 indicating most risky and 10 least, non-experts on average scored all risks below a 6. Experts, on the other hand, scored a third of the list (6 items) at 6 or above. This may suggest there is a more worried perception of technology, broadly, in non-expert risk judgments.

	Non-Expert		Expert	
Rank	Risk	Mean Rank	Risk	Mean Rank
1	Identity Theft	5.000	Job Loss	5.769
2	Account Breach	6.101	Account Breach	6.385
3	Job Loss	7.678	Identity Theft	6.577
4	Hacktivist Leak	7.980	Technology Divide	6.923
5	Auto-Drones	8.523	Bias Job Alg	7.192
6	Harassment	9.074	Discriminatory Crime Alg	7.231
7	Undisclosed third party	9.349	Hacktivist Leak	7.231
8	DDoS	9.403	Filter Bubble	7.654
9	Nuclear Reactor Meltdown	9.644	DDoS	8.269
10	Discriminatory Crime Alg	9.758	Undisclosed third party	8.462
11	Research w/o Consent	10.141	Harassment	9.346
12	Bias Job Alg	10.154	Auto-Drones	9.808
13	Driverless Car Malfunction	10.315	Research w/o Consent	11.154
14	Technology Divide	10.765	Nude Photos	12.038
15	Plane Crash	11.060	Driverless Car Malfunction	12.269
16	Filter Bubble	11.362	Nuclear Reactor Meltdown	14.308
17	Nude Photos	11.846	Plane Crash	14.654
18	Vaccine	12.846	Vaccine	15.731

Figure 5.1: Average comparative risk ranking by non-experts vs experts where items with significant differences (p¡.05 for two-tailed t-test) are highlighted.

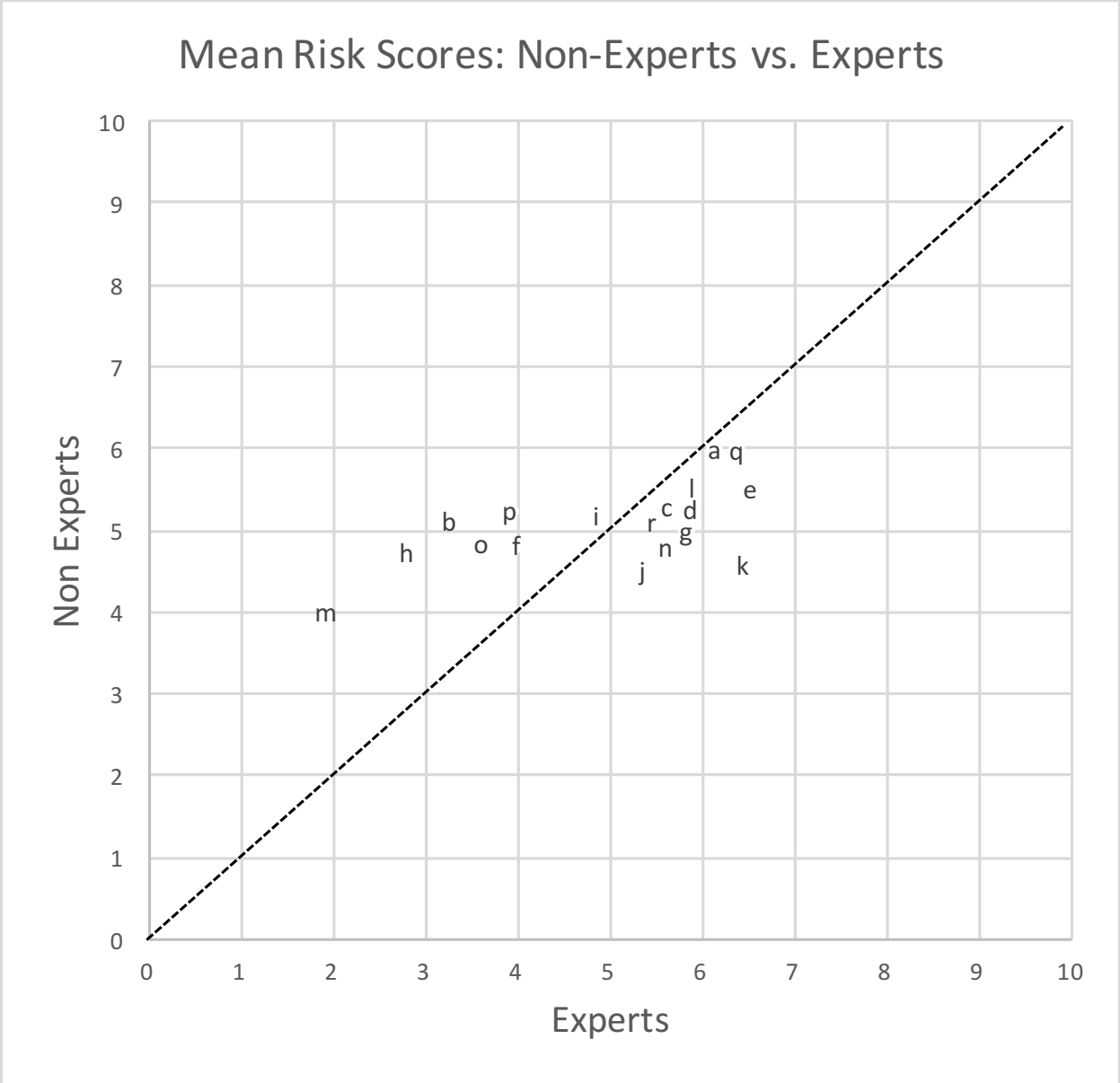
Though it is also possible that experts do not judge the broader risks of technology with enough gravity. Research without consent, while being only slightly different when comparatively ranked, showed a difference of 1.237 in mean score (p=.015) with non-experts ranking this as riskier than experts. This is not so surprising, though it is salient given the lively discussion currently ongoing around online research ethics [332].

5.2.3 Psychological Factors of Risk

We now move to examining how the psychological characteristics drawn from the original study played into perceived riskiness on the raw 1-to-10 scale. In our analysis, we were interested in the averages of factors that made up the psychological space surrounding a risk together with the individual correlations between a particular factor and its relationship to a risk score. Therefore, positive correlations reported below approximate the amount that a certain psychological factor influences an overall riskier perception of the associated item. Inversely, negative correlations indicate psychological factors that weigh in towards a lower perception of overall risk.

Our study broadly validated prior findings in the risk literature that show the more voluntary a risk is perceived to be, the less risky it is ranked and scored. For all except two items, higher perceived voluntariness correlated negatively with perceived riskiness. This is to be anticipated as, much like our acceptance of cars despite their danger, we expect people to take on more risk when they have chosen this risk and perceive themselves as in control. The two exceptions we saw made some sense – DDoS attacks and Death and Destruction from Autonomous Drones which, only for non-experts, voluntariness had a minor positive correlation with risk scores (r=.05 and .1, respectively). Given that there is really no voluntary dimension to either of these risks, it is unsurprising that voluntariness was basically uncorrelated even if slightly positive. If properly understood, the mild positive relationship may even be from a rightful recognition that owning insecure devices that allow DDoS to happen or not actively pushing to downgrade our military and suppress drone warfare are both voluntary choices that expand these risks.

However a novel complication from the prior risk literature is that both groups rated nearly



a. Identity Theft	g. DDoS	m. Vaccine
b. Nuclear Meltdown	h. Plane Crash	n. Tech Divide
c. Harassment	i. Auto-Drone	o. Research w/o consent
d. Discriminatory Crime Alg	j. Bias Job	p. Driverless Car Malfunction
e. Job Loss	k. Filter Bubble	q. Account Breach
f. Nude Photos	l. Hacktivist Leak	r. Undisclosed third party

Figure 5.2: Risk perception by experts vs non-experts for 18 technologies.

all risks related to emerging technologies as characteristically involuntary. That is, using our 1-7 Likert Scale, almost every risk carried an average score below 3 and in many cases below 2. This suggests that despite the consent processes built into most software and web services, the corresponding risks are not perceived as something being voluntarily assumed. There were a few minor exceptions where voluntariness ranked slightly higher. Within the non-expert grouping two risk items ended up with a mean voluntariness rating above 3: driverless car malfunction (3.43) and nude photographs (3.89) which were still below the middle of our scale. The expert list was the same – driverless cars (3.5) and nude photos (4.15) – with the addition of filter bubble (3.73). This is a good validation within our findings since among our items these are all risks that do have some immediate control by individuals – not getting in a driverless car, never letting nude photographs be digitized, and actively seeking news and information from sources outside of your ideological affiliations.

Other psychological factors of importance were perceived fear and severity, which both often carried a highly positive correlation with perceived risk. The perceived riskiness of destruction from autonomous drones, driverless car malfunctions, research without consent, and hacktivist leaks all had their strongest positive correlations with either fear or severity across both groups. This might imply that more than facts or reasoning some risks get a status due to the imagined consequences being frightening or intensely concerning. This may explain the somewhat surprising result that, even for experts, the perceived riskiness of driverless car malfunctions was most positively correlated with fear (r=.717), which outweighed the strongest negative correlating factor of the perceived understanding of domain experts (r=-.47). This suggests that the innate fear of the autonomous car crashing factors higher in one's risk calculation than their belief that experts are on top of debugging and testing.

Another notable result was that the expert group showed a positive correlation (r=.4) between how well domain experts understand the riskiness of hacktivist leaks–suggesting that the more experts understand the domain, the riskier it becomes. This is in contrast to most of the other technology risks, where more expert knowledge makes something less risky. This finding may imply

that as experts get better at understanding data breaches and leaks that they too could take part in hacktivist activities, or that even highly qualified experts cannot stop the ability for insiders and creative hackers to gain access to secrets. Other risks where expert understanding had a positive correlation (above r=.2) with perceived risk were discriminatory crime algorithms (r=.27) and filter bubble (r=.26). These risks probably warrant more niche domain expert opinions since they are both still new and ill understood by technologists broadly. It seems there is some belief that experts being on top of it might lower the risk, but it is not yet a strong one.

Figures 3 shows a graph that compares psychological factors between groups on two items that show significant differences between experts and non-experts: research without consent and filter bubble. As the graphs signify, experts appear to be generally more worried about the filter bubble whereas non-experts are more concerned about research without consent. Interestingly, experts have a fairly strong positive correlation between self understanding and perceived riskiness (r=.53) and similarly with likelihood of occurrence and perceived riskiness (r=.517). This finding could be straightforward, given that experts are probably more likely to have taken time to comprehend and break out of their own filter bubbles due to the current events and recent impacts from the issue. On the other hand, fear (experts r=.48; non-experts r=.295) and severity (experts r=.476; non-experts r=.269) have the highest positive correlations for both groups concerning research without consent. Though we cannot say definitively what exact fears they harbor or severe impacts they imagine, we will discuss in the next section some of the "worst case scenarios" that some participants stated with respect to these risks. Regardless, our findings do suggest that people believe there are real risks to allowing online data to flow into research without consent. Public perceptions of this domain could be heavily influenced by media explanations of recent controversies such as the Facebook emotional contagion study [224] and the public release of OKCupid data by a researcher [364].

5.2.4 Worst Case Scenarios

In addition to the psychological factors, asking participants about their idea of worst case scenarios for the technologies they consider riskiest gives us a more qualitative sense of what specific

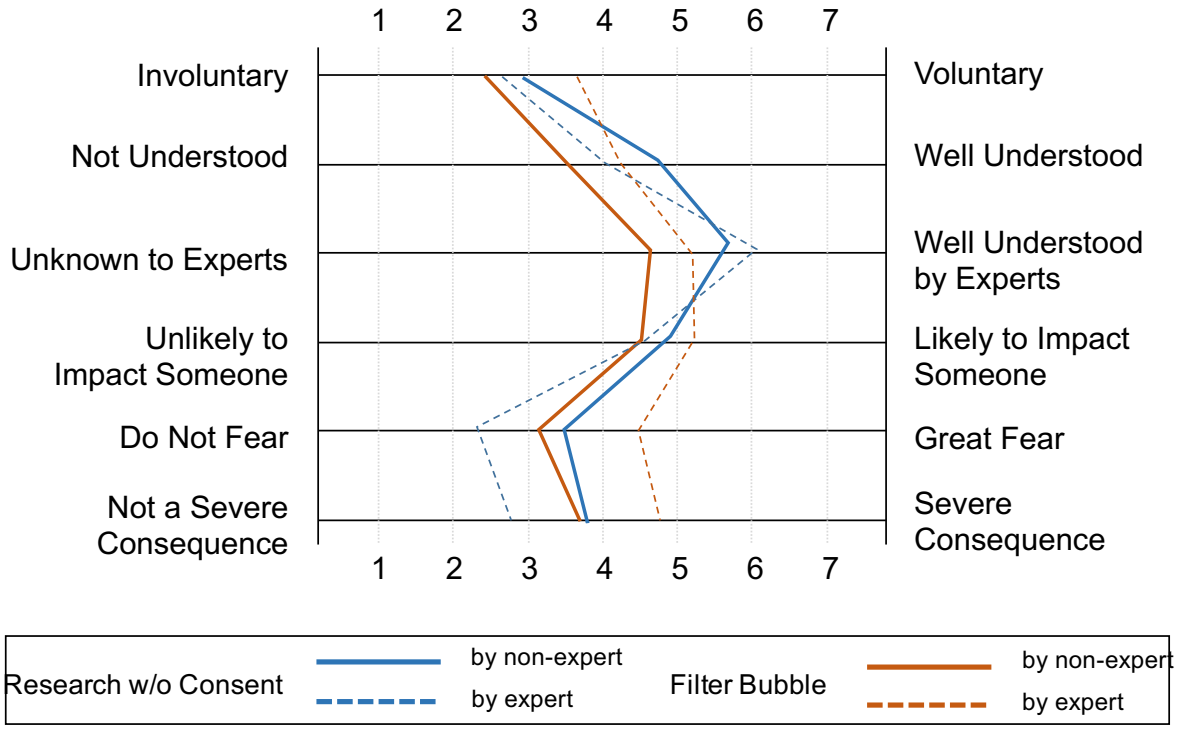


Figure 5.3: Comparing the psychological factors regarding the filter bubble (orange) vs research without consent (blue) as perceived by non-experts (dotted) vs by experts (solid).

Scenario Type	Risk	Example Quote
Financial	DDoS	"bank account drained of money"
Legal	Discriminatory Crime Algorithm	"being imprisoned for nothing"
Reputational	Nude Photos	"massive embarrassment and you lose your job"
Physical Harm or Death	Online Harrassment	"it slowly erodes you, so suicide would be the worst"
Societal	Hacktivist Data Leaks	"civil unrest making us vulnerable"

Table 5.1: Scenario Types, Risks, and Example Text from Respondents' Open-Ended Worst-Case Scenario Response

kinds of risks they consider to be the worst. Every technological risk was ranked within the top three by at least ten people.

We broadly identified five primary types of harm that we used to categorize the worst case scenario responses. These included: financial, legal, reputational, physical harm or death, and societal (see Table 1 for examples of each).

Based on the aggregate rankings of risk (see Figure 1), the top three technologies perceived as riskiest by both experts and non-experts are seen as financial and legal risks. Worst case scenarios for identity theft include "losing everything and being imprisoned" and "someone wipes out my entire checking account." Job loss from automation evokes fear of "widespread unemployment and the collapse of our economic system" and for account breach, "credit cards being used to make unauthorized purchases." Financial is also the most common type of perceived risk, also dominating responses biased algorithms, collection of public information, and data sold to third parties.

With respect to differences between experts and non-experts, technologies that experts rate as significantly riskier than non-experts (technology divide, biased job algorithms, discriminatory crime algorithms, and filter bubble) all have dominantly "societal" worst-case scenarios. For example, "bringing down critical services would cause mass chaos" (DDOS), "extremely polarized news and information to the point where no one is certain of truth" (filter bubble), and "extreme divides in economic opportunities and outcomes, leading to social collapse" (technology divide). One possible explanation for this finding is that more experts are able to foresee the broader impacts of

technology, than non-experts.

Another aspect of risk perception that these free responses highlights is that participants often have very different ideas of concrete risks for these technologies. For example, for DDOS, most participants mentioned worst case scenarios like bringing down emergency services or energy grids, but one wrote, "People being angry that they cant get on Facebook." Similarly, a number of participants, mostly among non-experts, thought that the risks around filter bubbles had to do with government controlled or censored Internet. For example, "They could choose what to let you know about what not to know so the earth could be under alien attack and they could just block that if they wanted to." The risk for online research without consent also showed a number of conflicting ideas of worst case scenarios, particularly among non-experts. Whereas some mentioned "invasion of privacy" and similar, which tracks to existing work around reactions to online research [363], we also saw what appears to be misunderstandings of what research of public data actually entails–for example, "getting lots of spam calls" and "could lead to stalking by the government."

5.2.5 Additional Risks

We also asked participants in an open response question what they saw as any other major technological risks to society. Less than a quarter of our participants included a response, which suggests that our list of risks was fairly comprehensive. Some of these were more specific or nuanced versions of risks we assessed–for example, "leakage of personal data through apps" or "hackers causing voter fraud." Others were technological risks not as related to computers or data–for example, "biotoxins" or "solar flares wiping out technology."

The theme that appeared most frequently among those who answered this question was addiction to or reliance on technology. Example answers were "young people of future generations will be lost without tech," "addition to technology," and "inability to get along without technology." In future work, this would be a good issue for further exploration, since it is clearly a risk that is on people's minds.

5.2.6 Discussion

5.2.6.1 Tradeoffs and Voluntariness

Without an understanding of risk perception, researchers and technologists could be missing critical information about decision-makers and users alike when making choices about regulation, technological design, or technology adoption. Our findings around voluntariness are particularly salient when it comes to user models of risk. It is a robust finding in risk literature that risks perceived as voluntary (such as driving and skiing) are seen as more acceptable to the public even if statistically they are dangerous [296]. When modeling user and consumer decisions around data-driven and networked technologies, a similar schema is often deployed to explain trading off privacy for convenience. There is an assumption that since users accept the terms of service and then provide information and content, any repercussions are part of the agreement and thus voluntarily assumed, despite a great deal of research having shown that it is unlikely that users read and understand terms of service [213, 175]. Similarly, releasing new features or adopting new algorithms that shape behavioral and social patterns are not considered problematic because use of the tool is voluntary. Our findings complicate this assumption since, in nearly all cases, technological risks are not seen as voluntary. This might illuminate the media addiction to reporting on technology's mishaps, since people may not really see these risks as something for which they signed up.

Given complementary findings around users' lacking understanding of the legal agreements to which they are bound, this perception of involuntariness might be expected despite common assumptions. However, this trend is actually more worrying since our expert participants generally ranked complex technological risks higher. This gap in expert and non-expert perceptions implies that users may not only feel they are being involuntarily harmed by the choices technologists make with data and features development, but that they may not even really realize what is at stake in the coming years. Taking this finding seriously, designers may want to offer more information and choices about how features may be affecting user experience. Technologists broadly could attempt to allow for more discussion, feedback, and disagreement around new features and data

practices, and perhaps be willing to hold off rolling out a feature when it raises more concerns than excitement.

Our exceptional cases to this finding actually further validate this take. Given that nude photos being shared without consent is seen as more voluntary than the other risks we examined, we can speculate that this perception is based on interpreting it as a voluntary choice to take or share nude photos at all. And while our results do not indicate it is perceived as fully voluntary, since the implication is the photos go beyond where they were meant to be seen, this does indicate an attitude that the burden is on the user not to digitize and share nude photographs. Similarly, with malfunctions in driverless cars, which may at first glance seem surprising to be viewed as a voluntary risk, we can speculate that this perception is based on seeing a choice to ride in a driverless car.

5.2.6.2 Most Technologies Are Riskier Than You Think

We also found that experts generally rank technological risks as comparatively more risky than non-experts. One takeaway we suggest from this finding is that it would be in society's best interest for researchers to focus more acutely on the complexity of education and public communications. With algorithms and AI mediating more of our relationships and institutions, yet being highly esoteric topics, it is concerning that the risks of these changes are perceived as stronger by the group who is involved in the creation of those risks. As we have already seen with issues such as climate change, vaccines, and immigration, once complex problems come to the attention of large swaths of the public who do not understand them, it can catalyze division, exploitation, and people acting out of the best interest of our society.

Some of these risks are also important for non-experts to understand as we move into new futures suggested by technology–for example, as AI leads to the automation of more tasks and potential job loss. It is notable that both groups rated job loss in their top 3 risks; yet, it is also interesting that experts rated this #1 versus non-experts #3. If there were to be a mass shift in jobs, e.g., from autonomous trucking and shipping, it is particularly important that this change is

well communicated and comes with a policy plan for mitigation. Otherwise, we may be steering toward another crisis emphasizing lack of trust between experts and the public. The potential for such a problem expands when considering that our findings suggest that non-experts are not taking the possibility of bias or discriminatory algorithms as seriously as are experts. Following workshops such as FATML (Fair, Accountable, and Transparent Machine Learning), technologists should take a stance on what makes and algorithm fair and communicate these standards to their users. Designers may also strive to make users aware of when they are being acted on by an algorithm and promote different kinds of reflection when displaying results.

5.2.6.3 What Makes People Afraid?

Our psychological factor analysis revealed that, besides voluntariness, fear and severity play heavily into people's perception of risk (both expert and non-experts). This is a challenging result to interpret without a deeper depiction of the imagined consequences. We suspect that fear has some interaction with voluntariness, as we often think of fears as having a somewhat irrational component to them as they are personal and can be uncontrollable. Such an explanation might help us understand why the risk of malfunctioning driverless cars was highly relative to the person's fear regardless of expert understanding.

Severity is even harder to analyze as this relates to the specifics of the imagined consequence, especially since some consequences even experts could not fully comprehend yet. We do know from our analysis of free response answers regarding worst case scenarios, that people's imagined consequences can vary considerably. For example, research without consent carried with it differing ideas of what this might mean–some non-experts think that it could relate to receiving spam or to government surveillance, which suggests a misunderstanding of what research on public data entails. Others may have a better idea of potential consequences, imagining that they get pooled and identified in categories to which they do not believe they should belong. Fear could also relate to some deeper belief about what it means to be studied. Regardless, the perceived risk around this practice, whether rational or not, is something that researchers should keep in mind when

collecting data.

With respect to experts' perceived fear, another finding that supports our understanding of current events is the significant difference in how experts versus non-experts see the filter bubble. Not only did experts rank the issue as comparatively more risky, but also many of their judgments of psychological factors were quite different. As public opinion and the media often blame fake news and political polarization on technologists, it is important that we get a better sense of how this problem is understood by the public. One might wonder if the filter bubble is truly just perceived as less risky by non-experts or if it is another risk that is poorly understood due to its technical provenance.

Our analysis of worst case scenario responses suggests that it could be the latter, since a number of participants framed the problem as being one of purposeful government control of the Internet. However, regardless of causality, it is also concerning that beliefs about self understanding of the problem correlated highly with experts believing it was risky. This suggests that the underlying problem may be situated in an educational space. As technologists deliver new personalized features to users, we may only amplify the problems surrounding the filter bubble without corresponding communications around how to work around or see outside of personalized experiences. Otherwise, given our results, it may be the case that people do not fully see how separate the lived realities of people across the country and world truly are and thus this problem is not taken seriously. Given other research that suggests, when asked, people are often skeptical of personalization [15], the fact that people do not see filter bubbles as a high risk may suggest a new challenge for designers: how to make transparent when and how personalized filtering is happening, and what the consequences are.

5.2.7 Risk-Sensitive Design

Taken together, our results suggest a number of thoughtful approaches to public communications, policy, and technology design. In addition to the implications already discussed, we also propose a design principle that we term—**risk-sensitive design**. Risk-sensitive design recom-

mends that design decisions regarding risk mitigation features for a particular technology should be sensitive to the difference between the public's perceived risk and the "acceptable marginal perceived risk" at that risk level. One must remember that technologies and risks are not always one-to-one and it may be a particular system or design that is creating a negative consequence rather than an individual technology.

We developed a graphical method as a tool to illustrate this design principle. Figure 5.2 plots risk perception by experts vs non-experts for the 18 types of risks we studied. Figure 5.4 is an abstract version of the same plot, that we will use to explain risk sensitive design. Each circle represents a risk presented by a technology for which its perception level (how risky each population perceives the risk to be) has been assessed. On the diagonal line, experts and non-experts perceive risk the same way. For brevity, we will refer to this line of equal risk perception as E.

For example, risk a is perceived by non-experts as no more and no less risky, compared to experts. Above the diagonal line, non-experts perceive risks higher than do experts, such as technology d. Below the diagonal line, non-experts perceive risks lower than do experts, such as risks b and c. The downward bending curve represents the concept of "acceptable marginal perceived risk." For brevity, we will use M to denote this curve. We introduce this curve in order to push for an important design heuristic—"how much can the public underestimate a risk before it is deemed unacceptable?" The higher the risk, the smaller the margin should be, which is modeled by the downward bend.

Observe that the two curves E and M divide the graph into three regions. Risk-sensitive design argues for different design recommendations based on the region in which a technology lies. We discuss each region in turn.

Most important, in our opinion, is the region represented by risk b. Here, the non-expert public grossly underestimates the risk b associated to some technology. This is reflected by its position underneath the acceptable marginal perceived risk curve (M). As shown, the gap between M and E is large and may be instigate broader harms than can be mitigated by simply refining these features. We take a provocative stance to recommend the underlying technology or technical ar-

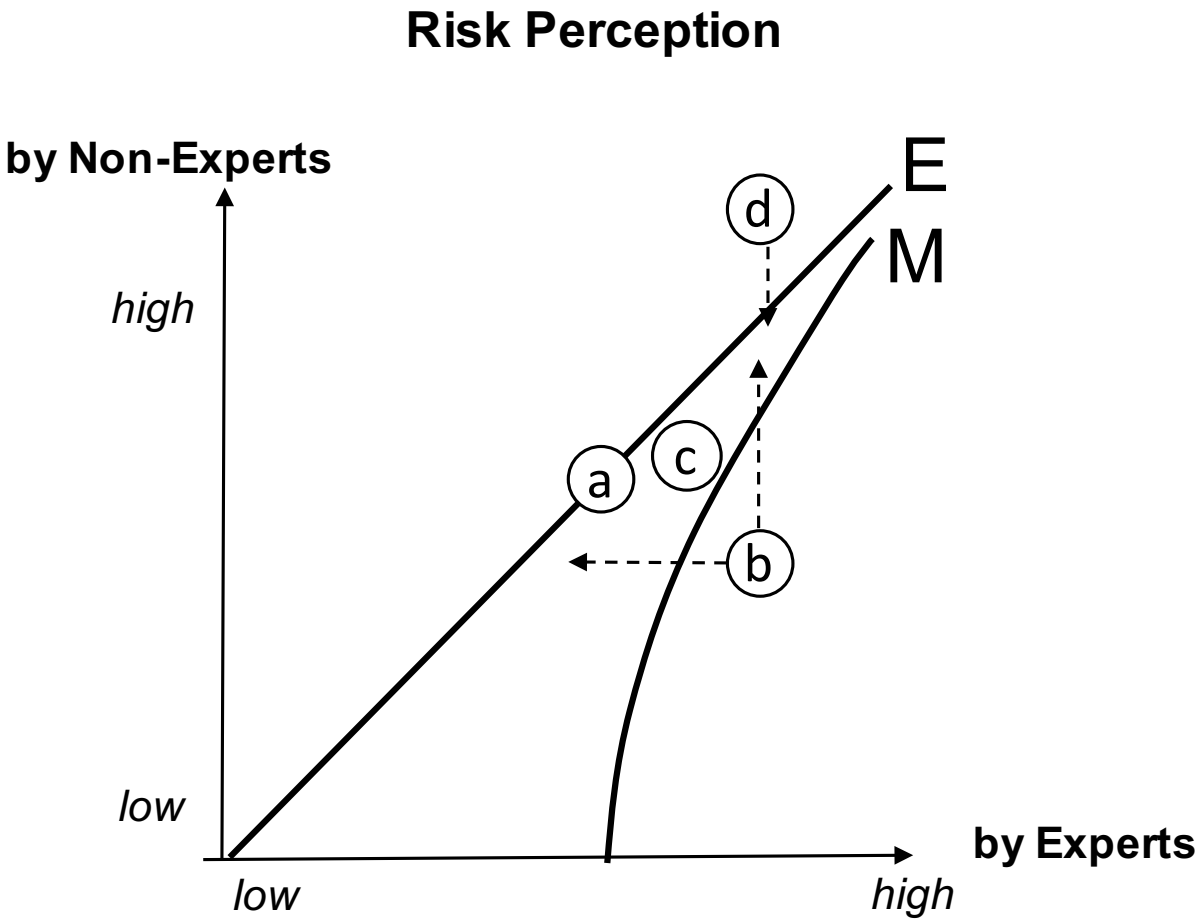


Figure 5.4: Risk-Sensitive Design proposes that risk mitigation strategies should be informed by the difference between public and experts' risk perception and the degree to which such difference is acceptable $(E - M)$.

rangement should be **reconsidered immediately**. Meaning, studying the impacts the technology is having, communicating to users the concerns related to the technology, working in consortium with others to elevate awareness and adopt policies toward mitigation, and potentially even recalling or scaling back the technology until the risk is better understood. Right now, given our findings, personalization technologies applied to content feeds relevant to public decision-making, which in turn cause filter bubbles, may lie in this region of concern.

We believe keeping a technology found to be creating risks in this region out there with no change or revision could be doing more harm than good to the society. While a technology is under review or even temporarily recalled, there are two potential actions. First, scientists and engineers can further innovate to reduce the risk exposure b caused by the technology, which would then be reflected in the graph by a leftward movement. Second, more investment can be made to increase public awareness of risk b, which is reflected in the graph by an upward movement. Both offer a path to exit this questionable region and enter the relatively safer region where technology c is. Once in a safer region, the associated technology can be re-introduced in its new form. We discuss this relatively safer region next.

The region flanked by M and E is considered relatively safer, but not entirely safe. Technology c's risk is still underestimated by the non-expert public. However, the gap is small enough to be within an acceptable range (above M). We recommend **strong safety and risk mitigation measures**. Depending on the kind of risk, examples might include two-factor authentication, compulsory user safety training, or changes to business practices or design. Additionally, better communication or education could also help reduce this gap. In some cases, tactics such as showing real-world examples of harm could be desirable.

The region above E is where a technology, such as d, is overly perceived as risky by the non-expert public. There is less concern that a person could be harmed as a result of not being cautious enough or misinformed. However, this situation could lead to diminished use of the technology or over-reactions when complications do occur. We recommend **adopting a communication strategy focusing on reducing fear and misconceptions**. The aim is to reduce the perception

of risk d, as represented by the downward arrow.

This paper contributes knowledge of where on the plot the 18 types of (mostly technical) risk lie. This paper further argues that a line ought be drawn for the acceptable marginal perceived risk (M) and offers design recommendations based on this line. But as to **where** this line should be drawn, we believe it is a subject of further public discourse, which we hope this study and model provoke.

5.2.7.1 Decision-Making with Risk-Sensitive Design

At the time we formulated the principle of risk-sensitive design, we had not yet completed our survey and did not know which technology risks would fall into the lower region where we would recommend rethinking the technology. Through our analysis, the two risks that fell closest to the lower-right region are filter bubbles caused by personalized recommendation systems and job loss caused by automation and AI. Both of these are pervasive and intertwined with many other benefits industry and research are likely less inclined to put on hold, or that are challenging to rethink. Deciding where to draw these lines is a difficult task, and proposing that certain technologies might be unacceptable within the bounds of risk-sensitive design is likely to generate debate.

However, we feel that more important than using this tool to make concrete judgments about a particular technology, is using it to provoke the right set of questions. For example, our qualitative findings show that filter bubble is a technology that is poorly understood by our participants. This suggests that, though experts rank it as risky, the gap might be more easily closed not by attempting to make the technology less risky but by educating the public about how it works and what the risks are. With respect to job loss from automation, this risk is co-evolving with the continued advancement of technology, and should be approached with a combination of risk mitigation in design and public communication and education to reduce the expert/non-expert gap. Beyond design, this finding implicates the need for policy considerations.

Risk-sensitive design calls for designers of new technologies to go beyond traditional risk

and benefit analysis and to also pay attention to how their users may perceive risk, provoking thoughtfulness about how to introduce a technology in a socially responsible way. One should bear in mind the question of "where might risks created by my technology enter into this space" and make efforts not to prematurely introduce a technology if evidence projects an entry point below M in our graphical tool.

We also recommend that risk perception should be included as a factor to design as early in the design process as possible, rather than as an afterthought once a technology is already built and deployed. For example, focus groups and field observations in formative studies should include risk perception as a factor, not just an objective risk assessment. Summative evaluation should also include instruments to measure risk perceptions with study participants, and multiple expert opinions should be weighed as well. The survey instrument we used is one model for obtaining risk perception data.

One limitation of this approach is that we assume risk is known and can be predicted by experts, as a measurement alongside user perceptions. For some technologies, however, risk is not known until it is introduced. Future expansion of our model could include uncertainties, such as modeling a technology's position in the graph as a probabilistic bubble rather than a dot, and movement within the space as a funnel rather than a line.

5.2.8 Conclusion

The above study applied an instrument from the risk perception literature to analyze thinking about emerging technologies. We found that generally users do not think of risks exposure from technology to be voluntary. There were also considerable differences in how risk is perceived by experts and the lay public, which may explain why certain problems, such as the filter bubble, have become so concerning. Our paper ends with a discussion of risk sensitive design, hoping to provoke continued conversation about how technologists should respond when there are large gaps in how the public and experts think about risk. We hope to see HCI researchers continue study in this area and advance the conversation further.

5.3 More Than a Show: Using Personalized Immersive Theater to Educate and Engage the Public in Technology Ethics

5.3.1 Prologue

This paper goes into depth on the biggest project within this dissertation work – Quantified Self. Written as a theater piece where the audience was able to play with technology, talk with actors and one another, the work represents a robust **narrative template** intentionally designed for research purposes. The project took nearly a year to produce and engaged hundreds of public participants from all walks of life. The goal was to create a space that would support broad public dialogue about ethical issues around data. Within the piece, the different characters take on different lenses for thinking about the multiple moral perspectives that may be taken on the enterprise of Big Data technologies. The audience was surveyed on their opinions and attitudes related to the show and its topics. This is an example of how well-devised narrative can create incredible opportunity for public engagement around ethical issues in technology. One of our primary findings that will be discussed was the audience's interest in attending more events and programming to get further educated on the topics following attending the show.

5.3.2 Introduction

Choices being made about how technology is designed, developed, and used shape the character of public and private life. Yet, there are still many open questions about the purpose of technology in human life and society: What should and should not be done with our personal data? [29, 301, 90] How do we apply laws and protect rights in the space of technology? [27, 197] To what extent should we allow choices by technologists to disrupt social norms? [123, 29, 160] In this techno-society, it has become imperative that we broaden participation in the discussions around technology development to support a negotiation on how to build a future that works for everyone; not just a small elite group of technologists and policy makers.

Creating an opportunity for this broad discussion is truly difficult. Not only does it demand

engaging diverse groups to participate in discussion and reflection, but also one must communicate complex issues in a format accessible to non-experts. To date, perhaps the most successful format for wide engagement in technological and scientific problems is to represent them in artistic and diegetic manners [188, 244, 189, 233, 44, 218, 219]–e.g., sci-fi, theater, and concept art. Researchers too use design fiction [58, 307, 313], design ethnography [210], user enactments [248, 124], and other forms of speculative art [118] to pose questions about the future. It is within this trend of combining art and technology to raise opportunities for discourse and research that inspired our project, "Quantified Self: Immersive Theater and Data Experience."

Our goal was to design a public engagement program that raised awareness and discussion around what it means for mass amounts of personal data to be owned and used by third parties. The motivation was to go beyond questions of privacy and focus on what it may be like to live in a world where all this data is used. Will people like that reality? Will it be fair? What control should we have over our own information?

Inspired by recent HCI work on enactments and design fiction and the gripping power of contemporary immersive theater [66], we embarked on a year-long production that would engage, educate, and provoke dialogue about the future of data. The result was Quantified Self which drew over 240 audience members of diverse backgrounds to six performances and was co-produced by a cross-disciplinary team of 22 students representing computer science, electrical engineering, theater, fine arts, and physics. Using a heuristic to frame our design considerations, Quantified Self differentiated itself from past efforts in that audiences both experienced and participated in a designed fictional world with both theatrical and technological (personal data-driven) elements.

5.3.3 Related Work

5.3.3.1 Why We Need More Conversations about Our Future

It should come as no surprise to HCI researchers that there is a desperate need to educate and converse with the public on the nature and potential futures of technology, in particular the uses of

Big Data. Recent events and research raise red flags that all the new ways of using data may not lead to shared prosperity. The groundbreaking work done at Pro Publica highlighted how machines are capable of being discriminatory much like humans [27]. Further work has made this risk evident as police face databases are biased toward African American faces [201], word embeddings have proven to encode gender stereotypes [62], and even behavioral advertising is codifying unfortunate racial differences [312].

There has further been attention payed to the new powers of data through predictive inference. 2016 in particular made evident the reality of algorithms shaping our society [123] and affecting public discourse [90]. And while prescient members research community may have identified warning signs [321, 94, 65, 250] for problems these changes may create, it's unclear technologists are really connected to the best interests of the public and their users [160].

Past research has shown that the mass move toward personalization is seen as unwanted or even worrying by some users [320, 29]. Even studies done by Google engineers availed that users do not see themselves as receiving the benefit of mass data acquisitions and sharing [55]. Moreover, we should not be surprised by the umbrage taken by certain communities [301] when plenty of HCI research points to rampant misunderstandings around the nature of the terms of service that legally bind us and the underlying technologies that track, categorize, and target us [223, 170, 213].

5.3.3.2 Approaches to Studying Problems of the Future

While we have come to understand the problems of ignorance and public concern with technology, researchers have continually worked to discover our best ways to communicate and study the problems. There is a long history that suggests sci-fi and other forms of speculative art are critical vehicles for reaching the broader public and raising awareness about the problems the future may bring [118, 188, 331, 61, 33, 291]. This history of interaction between technology and art has given rise to an appreciation of the influence and impact by researchers [219, 220, 208, 233, 211]. Knowing the affordances of art for technological and scientific discourse, Bruce Sterling formalized the idea of design fiction in order to establish creative methods for researchers to employ imagination

and pose questions about technology where social, political, and emotional content is integrated [307, 58].

Years later, we see a rich array of approaches being explored to utilize art as a means toward critique [118, 33] and conceptualizing where certain areas are headed [115]. Though not necessarily employing art, future studies is an area that has attempted to formalize the study of possible futures [217] and the potentials for utopia and dystopia [67, 98]. However, within the space of art and design, many incredible projects have emerged that have affordances for research.

The notion of a design or anticipatory ethnography [210] has brought qualitative methods closer to art and artists, assessing details about worlds that do not yet exist and the thought that goes into making them. Most influential on our thinking has been the creative approach of integrating theatrics into the study of how people feel about the future. Enactments [248, 124] and 'lived informatics' [125] allow us to take scenarios and design prototypes a step further to construct sites of research [365]. That is, by placing people into environments and scenes representing possible futures, we can leverage the role of improvisation to see how people act and feel within these settings.

It is in the overlap of these areas–design fiction and user enactments–we found potential for a project that can at once engage and educate while also allowing for discussion and listening. Thus, it was our goal to meld together some of these pioneering methods to push the potential for public engagement and research regarding the future of data use and ownership; specifically, how it is shaping public and private life.

5.3.4 Design Space

We begin by offering a heuristic for thinking about the design space related to the problems discussed above: informing the public, engaging a wide audience, provoking dialogue, and learning from our participants. The heuristic was designed to simplify thinking about other tech/art projects that were already achieving educational goals while building opportunity for other kinds of discovery such as research, inter-group conversation, and heightened individual awareness of technology's impacts.

The dimensions we sought to map were: 1) The degree to which art is being used to represent human-centric problems such as emotions, relationships, and qualifying experience. 2) The degree to which technology is being used to represent quantitative content such as mined data, live sensing, and mathematical simulation. 3) The degree to which the overall content is fixed prior to engaging an audience. 4) The degree to which content is improvised and adapted, allowing audiences to inject agency into the final experience.

As you can see in Figure 5.5, we began mapping projects on this 2D axis where extension into each quadrant brings with it specific affordances. What we were going for was a general heuristic that could represent the intersection of technology and art at a broad scale. While we are focused on the possibilities of highly cross-disciplinary endeavors such as design fictions, this graphic can even be used to understand simpler interactions between technology and art such an interactive graph. One may even think of a GUI as being the application of a bit of art to technical content to make raw memory and data meaningful to humans.

Breaking down each quadrant, we summarized the unique affordances of each part of the space. Art is known for its profound ability to express human problems and breed empathy in its expression. Modern technology, on the other hand, is a set of tools where using computers provides us with powerful ways of representing quantitative problems that go beyond the limitations of our bodies and manifest physical reality. Adding the other dimension, improvised vs. fixed content, the use of improvisation within technology is often our attempt to customize interaction and promote user-guided exploration, a common HCI problem. Similarly within art, the utility of improvisation is to allow audiences to have agency and co-create artistic experiences that are not one-way transmissions. However, of course, sometimes one way transmissions in terms of a static graph or a fixed story can have edifying purposes themselves.

Using this mapping, we can represent a variety of tech/art projects within HCI research and beyond. Figures 5.6.a to 5.6.f show our heuristic applied to other projects. We see traditional theater primarily occupies quadrant 1 (Figure 5.6.a), being a fixed form of art, whereas design fiction begins to bridge the imaginings of art with a focus on how technical problems might be

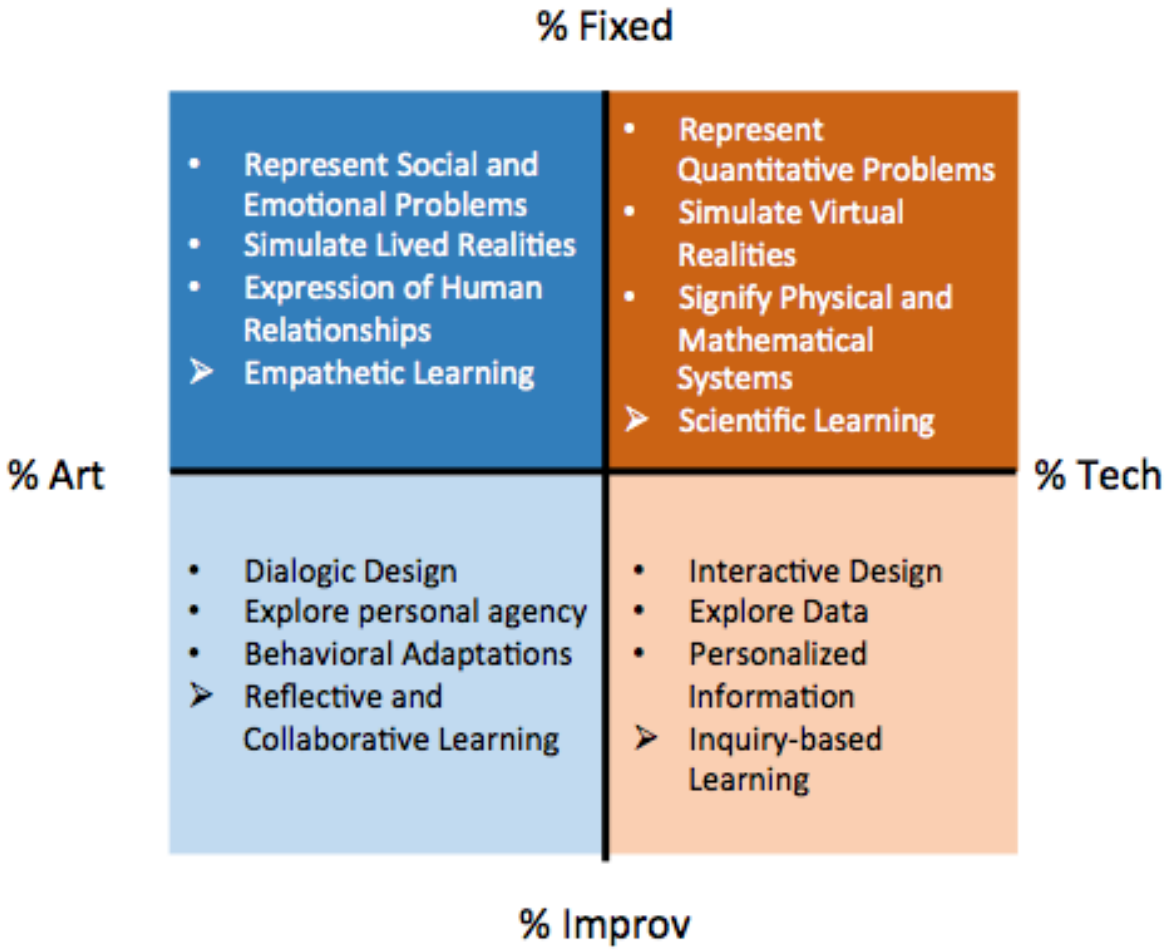


Figure 5.5: Our graphical model of the affordances offered by a tech/art project's design structure.

solved. We further see how interactive InfoViz (Figure 5.6.e) is meant to leverage art to make patterns in data recognizable, but also leverages interactivity to promote some exploration and discovery unique to the user.

After mapping out a range of prior efforts, we developed an excitement for possibilities emerging from the design space of enactments. Enactments use improvisation with both art and technology to allow users to explore designed future-oriented scenarios and in turn research their reactions, attitudes, and beliefs. However, we wanted to balance the fixed and improvisational content more, occupying a central position in the map. As illustrated in Figure 5.6.1, the goal was a perfect balance between technology vs art and between scripted vs improvised content. Our belief was that achieving this balance would allow us to maximize our four goals for a public engagement project: engage, educate, discuss, and learn. In short, we hoped to use the affordances of art to engage audiences on the topic while the technology offered real technical artifacts for our environment. Our fixed content was meant to educate the audience on some of the foundational considerations regarding technology and data ethics, and our improvisational content to provide the audience with agency to discuss and explore.

5.3.5 Approach and Methods

To obtain a balanced experience as laid out above, we developed "Quantified Self": an immersive theater piece where audience members were part of the performance. They were able to walk freely through the set, converse with actors and other audience members, and interact with the set and story using their own volition and at their own pace. Technology exhibits, or "companions" in our story line, were installed throughout the set, offering personalized interactions using social media data that our audience shared at ticketing.

5.3.5.1 High-Level Design

Beyond the artistic presentation, to further increase engagement, we aimed to maximize interactivity and personalization. Much like our contemporary technologies, we wanted a theater

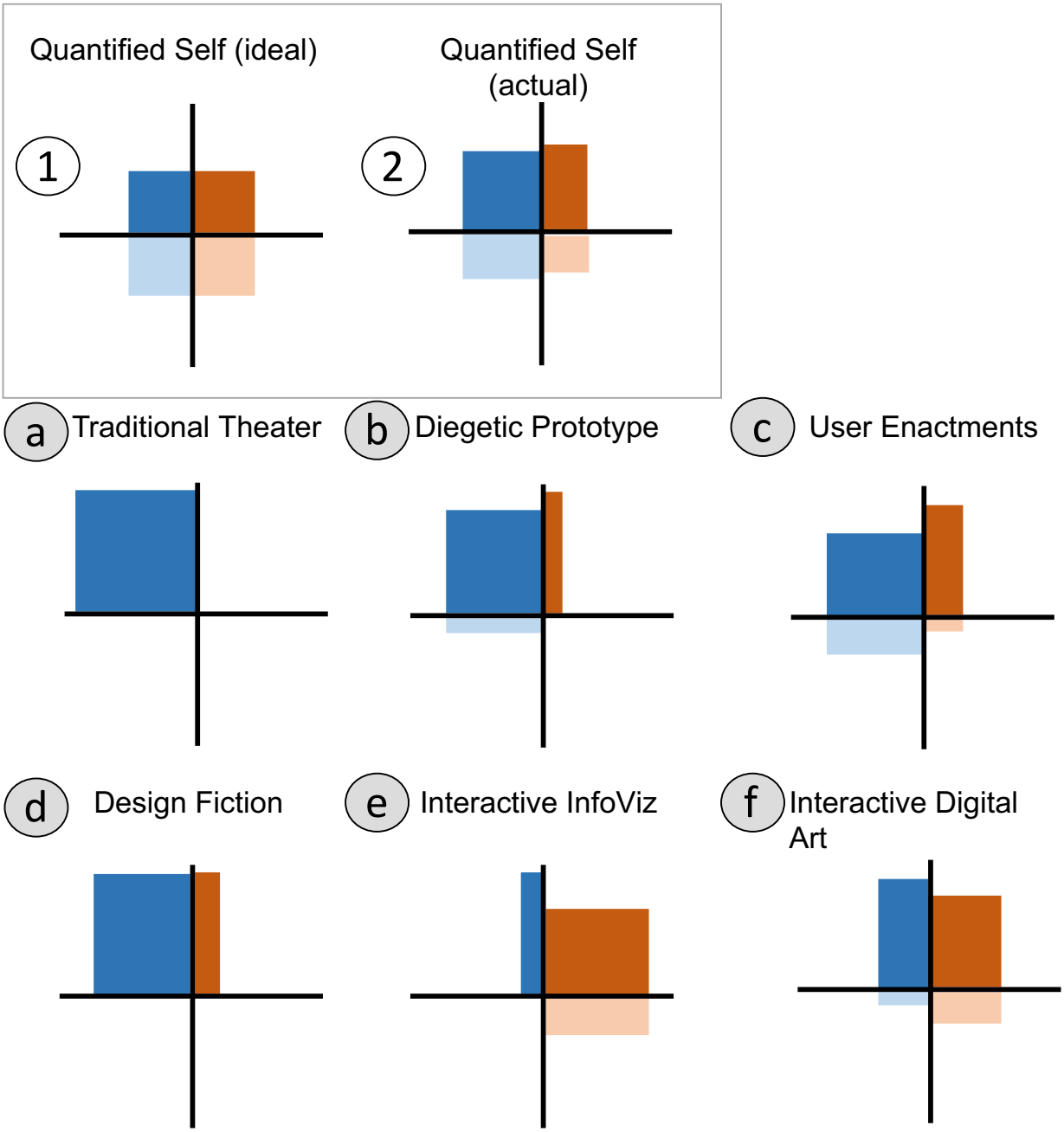


Figure 5.6: The goal of Quantified Self is to afford a balanced experience across all dimensions of our heuristic (1). We came close to meeting this goal (2). Below are examples of other projects using our heuristic for considering the affordances of different designs (a-f).

performance that adapted to the audience. Designing an open format also promoted discussion by creating communication pathways between the different communities coming together. Audience members could talk to friends, to strangers (often from a different background), actors, and, during the talk back, to our production team. Different perspectives in these conversations were embedded into the story and could happen organically by attracting a broad audience.

In terms of content, the show incorporated four primary educative themes on technology and data ethics:

Approximation of Self (A) Our goal with this issue was to discuss the controversial idea that behavioral metrics and online presentations are enough to capture the nature of a person. Of course, in actuality, human attributes quantified and aggregated online are approximations using statistical methodologies. However, the results of these methods have serious consequences on the shape of online and offline experiences [254]. In what cases is it ethical for them to be used to make important decisions about a person or to schedule and shape their activities?

Data Ownership and Privacy (O) Data is primarily owned by the companies providing the service, leaving the user with little control over how their data is used. Where law could be a safeguard, we often find that people are blindly clicking and agreeing to contractual agreements they largely do not understand [40, 77, 170]. When browsing websites, tracking can occur without opt-in consent. Companies offer free services by being able to buy and sell these everyday interactions. What should companies have the right to do with our data? What transparency should there be regarding these uses?

Presentation of Data (R) This issue looked at how people may be influenced or manipulated by the presentation of information that seems plausible but may be fallacious [338]. How does the presentation of data impact how you interpret it? Can showing quantified results breed undeserved trust?

Personalization (P) The issue of personalization wrestles with the current trend in technology

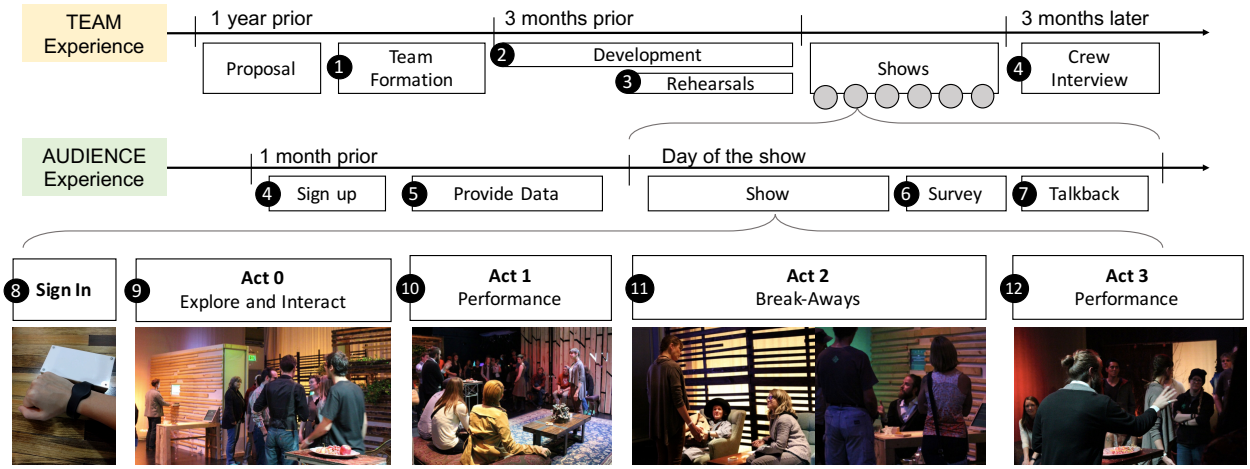


Figure 5.7: Time line for Quantified Self production.

that values interfaces and systems that use current and past information to tailor an experience to a particular user. Critically, there are times where personalization may create privacy concerns by exposing or implying information about someone or be an undesired hindrance such as when experiencing a search bubble phenomenon [320, 123]. How much of our world should be personalized? Would we like a highly personalized world?

5.3.5.2 Design Process

Quantified Self was developed through an iterative design process over a one-year period (see Figure 5.7). The production lead developed the initial idea and storyline, setting the theme and goals. The co-producers worked with the lead to develop the sign-up process, technology back-end, and the initial script. Three months prior to the show, the broader team was finalized (Figure 5.7.1).

To promote diverse discussions within the show, we had a team that crossed a breadth of technological and artistic backgrounds to support in the development and implementation. Our production team primarily consisted of 22 university students, crossing 7 different departments, with additional support from industry partners and faculty advisors (Figure 5.7.1). The production was led by a third year PhD student in computer science and co-produced with two other PhDs (one studying theater; the other interdisciplinary technology) and an industry data scientist. The leads

were supported by 19 other students (15 undergraduate, 4 graduate-level) divided into two main teams: technical and theatrical.

Undergraduate students were brought on the technical team to design and develop exhibits. The complete theatrical team started rehearsals and helped develop the characters through practiced improv. Overall, there were 5 rehearsal days (Figure 5.7.3) where the technical team worked directly with the theatrical team, sharing technical knowledge with the actors and receiving feedback on their exhibit designs.

5.3.5.3 Building The Experience

Script

The first piece of the show developed was our script, which was written by the project lead. Here we look at the two primary considerations that went into writing the script: the perspectives it embedded and how it supported the desired experience design.

The show had an overarching narrative following an ethical conflict within a famous tech company, DesignCraft. Immediately upon signing up for the show, participants were invited to a party for their supposed friend, Amelia, who was a star employee at DesignCraft. As the story unravels, they learn that Amelia is an experimental AI created using their personal data, who, herself, has begun grappling with the ethics of how the company uses her and its vast trove of data.

Within this broader plot arc, main characters were written to offer contrasting perspectives on our issues. Don, the CEO of DesignCraft, represented a business and innovation perspective. Lily, the chief data scientist of DesignCraft, held scientific and humanitarian views on the possibilities of Big Data while struggling with some privacy concerns. Felicia, an ex-DesignCraft employee, offered a critical lens of technology infiltrating and destroying the best parts of human relations. Evan, a hacker, saw technology as an opportunity for exploitation and intended to similarly use it to exploit DesignCraft. Amelia, a humanoid AI, struggled with the idea of being merely an instrument for technology and the artificiality of knowing people only through data. Felicity, an FBI agent, believed data could support a more secure society. Bo, the chief marketing officer at

DesignCraft, felt strongly that technology was entertaining, useful, and enjoyable and was willing to make this trade-off for any privacy concern. Finally, Veronica, a reporter, was concerned about the politics and intentions of the companies working with everyone's personal data.

Throughout the experience, participants were free to observe and follow different elements of the story, along with potentially discovering that they were actually involved in certain parts of the story. Beyond observing, the script called for actors to engage the audience members in open dialogue through a list of questions and unscripted-yet-topical scenes. During the rehearsal period, actors received training in data issues to be able to have informed discussions with the audience.

Following our heuristic to balance improvisation with fixed structure, we divided the show into four acts, two of which had a unified performance and two of which had multiple scenes happening in different locations. In Act 0 (Figure 5.7.9), after signing in, audience members entered Amelia's apartment (the stage), where they could freely use the technology exhibits and talk with actors (who may or may not be revealed as such) and each other. This was followed by Act 1 (Figure 5.7.10), a scripted scene where the main characters set the narrative conflict of Amelia realizing she was an AI-based upon audience data. Act 2 (Figure 5.7.11), was a series of break-aways, where audience members were encouraged to follow a character to hear a different side of the story. The story culminated in Act 3 (Figure 5.7.12), with a final scripted dialogue.

A difficult factor in writing the script was deciding how to fit technology fit the world. Since the script was written prior to the creation of the technical artifacts, we hoped to leave open possibilities for iteration and not require technical feats that might have been unfeasible given the timeline and budget. In the end, two (of the final ten) technological artifacts were explicitly built into the script allowing the remaining artifacts to be malleable.

Exhibits

As participants signed up for a ticket (Figure 5.7.4), they had the option to share their personal data with the show (Figure 5.7.5). Upon arrival, they were checked in to the party and given a Biobracelet (an RFID wristband) (Figure 5.7.8) that allowed users to sign into each of the 12 digital "companions" within the set. "Companions" were the name given to the fictitious

product line DesignCraft created and provided the plot basis for the interactive exhibits. A digital companion is an interactive set piece which uses participants online data (e.g., Facebook, Google, Twitter) to create personalized games and experiences, aimed at demonstrating many possible uses of personal data from mimicking online dating to the use of data in a job interview.

Data

A major consideration for this show was how to receive and manage data. We began by working with a lawyer to develop a simple terms of service that promised the user that no one would have access to their data prior to its use in the show and that all data would be deleted immediately after the show. There was also a corresponding set of bullet points presented to the user during ticketing about how their data would be used. For this pilot iteration of the production, we chose to take no anonymous statistics or behavioral metrics from user interactions with the exhibits.

Working with major data providers (e.g., Google, Twitter, Instagram) was mostly painless. However, we did have a real problem in getting Facebook to approve our app since their approval process is not designed for any kind of offline artistic experience like this. In the end, we had to film ourselves interacting with the interfaces in a rehearsal setting to get ultimate approval.

In order to keep data safe, we strongly encrypted all personal information offered by the audience. This involved generating a key pair during the ticketing process. The public key was used to encrypt all data after the mining and processing was completed. The private keys were stored in a further encrypted database that had five associated private keys, given to our four project leads and the lead set designer. At show time, two of the five keys needed to be entered in order to unlock the private key database for that night's show.

Finally, each audience member was given an RFID bracelet upon entering the performance, after showing photo ID (Figure 5.7.8). The bracelet's encoded ID was associated with each user's private key, allowing each user to individually choose which exhibits to check into, unencrypting the required data, or not.

Set

Exhibit	Description	Data Issues	N	Note
Own Up (1)	A collaborative experience where users see anonymous online quotes and choose to own up or not.	A	2-4	
Meet Your Match (2)	A compatibility engine to determine whether another person is a romantic match	A	2	
Mirror, Mirror (3)	A mirror with personalized information and messages	P	1	private
No Application Required (4)	Data is used to determine qualifications and personality traits for a job	A	2	with actor
Memory Wall (5)	A physical data piece where users map the order of major life experiences	A	1-6	non-digital
Highly Recommended	An either/or game that slowly questions users' judgment if they deviate from their data	A	1	
Discover your Inner Desire	An ever more complex series of terms of service agreements	O	1	
Wellness Booth	A system to determine happiness levels and mental health state from personal data	A	1	private
You and Your Libido	A recommender for who you're most attracted to from Facebook friends based on porn preferences	A	1	private
Interpret the Truth	Determine what facts are true or false about the world and user	R	2	
Infovision	See how cookies track users based on likes and dislikes of the news	O	1	
In the News	Get a personalized DesignCraft news article targeted to a user	P	1	

Data Issues: P: Personalization, A: Approximation of Self, O: Data Ownership and Privacy, R: Presentation of Data

Table 5.2: Selected exhibits to expose audience to certain personal data use issues. N denotes the number of users an exhibit is designed for.

The set was designed to promote opportunities for different forms of interaction. Our layout had a central living/dining room where the main scenes occurred. We then had a game room with a bar (snacks only), a den, an art studio, a bedroom, and an office. Each room had at least one exhibit in it and the layout was meant to facilitate a mixture of private conversations while also letting people see what was going on and float between scenes. Most of our walls were made of recycled crate wood slats that allowed the audience to see through into other rooms. Besides the bedroom, which was designed to allow for closed-off conversations with Amelia, the rest of the set was designed so that the main living/dining area was visible from anywhere.

Talk back

A last feature of the experience was a talk back (Figure 5.7.7). Following each performance we brought our entire cast and behind-the-scenes crew out in front of the audience for a discussion. We would first ask those interested to take our survey (described below) and afterwards we would open up a dialogue with our cast and crew. This would start by our production lead explaining that the point of this piece was to build engagement around these issues and that now was their time to ask or discuss anything. From here onward, we would allow the audience to get into a queue to ask questions to our team and would leave the floor open between questions to allow questions and clarifications across audience members. After 20-30 minutes we would let the actors leave, but sometimes the conversation would continue for up to an hour.

5.3.5.4 Constraints

Given our design goals and production requirements, we found ourselves having to deal with a number of constraints in our implementation of the production. To begin with, our research methods were explicitly chosen to take a backseat to the goal of broadening participation and discussion. This meant the engagement value of the show was taken more seriously than devising hyper-specificity in the dialogue or exhibits for research of particular technologies or systems. It also meant that we wanted our survey to be fairly short since we were already asking for 2 hours of attention at the show and did not want to limit our study to those willing to do a lengthy research

program afterwards.

Further, while we were interested in collecting anonymous statistics and taking observational notes, we also did not want our audience to feel overly exposed, being observed at every juncture. In an informal pre-survey, we found that potential participants were hesitant to share their social media for research. We wanted our terms of service to have an extreme favoring of user rights to promote trust and allow people to act candidly during the performance. The research was also meant to be optional, making the possibility of monitoring actions during the show difficult.

These research aims along with other production goals were most constrained by timeline. Restrictions of our funding timeline and the academic calendar combined to create a very intense schedule. This meant the script had to be finalized before we developed multiple endings (something the production team all believed would serve our improvisation goal well) and we were not able to do much in terms of pre-show audience engagements.

5.3.5.5 Research Methods

In consultation with our IRB office and in light of the constraints to achieve our engagement goals, we adopted two primary forms of data collection aimed at the audience and the production team.

To determine impact on audience members, we implemented a post-show survey (Figure 5.7.6). We ran six performances of Quantified Self over four days, each with 40 ticketed slots, and an approximate total attendance 240 guests. After each performance, we invited participants in the audience to stay for a 15-minute anonymous survey. The survey consisted of 20 questions aimed at obtaining information on demographics, prior experience with data and theater, general attitudes towards data ethics, and reactions to the performance. The majority of the questions had participants rank statements on a Likert scale. In total, 179 filled out the survey, although several surveys left a section or question incomplete (as reflected in the differing numbers in the results).

Utilizing a post-survey afforded us greater participation and more accurate audience experiences. First, it allowed for participation in the production by those who were weary of sharing

personal data. Given the hesitance and uncertainty around who would look at the personal data that was shared, we adopted a policy where none of that personal data would be used. Second, since research recruitment started after the performance, audience members experiences within the production were genuine. Without the added layer of being in a research study, they were unimpaired by the sense of being watched or recorded beyond the semi-public nature of the show. Third, given our own moral grappling with data ownership, we felt it most ethical to give audience members full control of what information they would like to share with us.

In addition to our audience evaluation, we documented the design process and performed post-production interviews with the crew to look at the educational impact and reaction to being a member of the production team (Figure 5.7.4). We performed 15 crew interviews (5 technical team members and 10 theatrical). Interviews lasted approximately 20 minutes. We focused questions on learning (both self-reported and content questions) and impacts of the show to topic understanding and behavioral changes.

5.3.6 Results

5.3.6.1 Show

The show itself was a success considering its scope and complexity. While we did ultimately see this as a pilot for a future scaled and improved effort, the reception clearly highlighted the potential of such a project. All six of our shows, which offered 40 slots each, were filled within the first week. We actually allowed in more than the 240 ticketed guests as each night we had a few extra guests come to the door and seek entry. Normally we would let them in as space allowed; though they could not interact with the technology. As the remainder of the results will show, the audience generally left wanting more. Every night there would be a portion of the audience who would stick around, talking, until the last allowable moment. Many technologists in the audience explicitly thanked us for engaging people on such important issues and the excitement it generated throughout the crowd was tangible. As a student-run performance, we received write-ups from

university and local new sources. Word-of-mouth spread afterwards, with the help of our granters, that we now have several theater companies waiting to talk about a follow-up run of the production.

5.3.6.2 Audience

5.3.6.3 Educational Goals

One of our design goals was to educate all participants in this effort, including the audience and the production crew. Regarding the audience, we report survey results relevant to their educational sentiments and perceived content accessibility. One question asked if they felt informed, confused, motivated to learn, uncertain, upset, hopeful, or concerned following Quantified Self (checking all that applied). 51% of audience members reported feeling more informed after the experience, while 16% reported feeling more confused and the remaining 33% reported neither. The number that reported neither are likely to have been from the number of people who attended who are already highly engaged in these topics.

While it supports our goals of inter-community conversation that some people came in more informed, we would hoped the percentage would be higher. However, an encouraging sign is that 62% of the audience members reported they were "motivated to learn." So, even if our show did not give everyone enough content to feel informed, it may have provided others with the impetus to go do their own research.

People leaving upset was something with which we were concerned. The scripted ending of the show had our protagonist, the humanoid AI, killed (shutdown). This pessimistic ending, we worried, may have blurred the more balanced conversation throughout; however only 14% left feeling upset as a result.

Another question asked audience members to compare the show to other more common modes of getting educated. Overall, we found that on average our audience somewhat agreed our show was more accessible than reading an article ($3.53/5, \sigma = 1.53$), taking a class ($3.46/5, \sigma = 0.98$), and researching online ($3.52/5, \sigma = 1.02$). We also included an unlikely alternative—reading a

privacy policy. Not surprisingly, the audience most strongly agreed the show was more accessible ($4.33/5, \sigma = 0.77$).

One limitation of our finding here is that we only asked the audience members to compare to common alternatives as opposed to novel experiences they may have in the past. We are also limited to sentiments rather than, say, subject-matter expertise developed. Because we could not do a pre-survey, any content-based learning would have been impossible to assess. Given the size of the audience, it would also have been difficult to require lengthy or difficult pre/post instruments. In retrospect, adding a pre/post focus group may been the right balance.

5.3.6.4 Engagement

One primary goal of this show was to engage a broader audience in conversations of ethics. In particular, we aimed to start a dialogue between technologists and non-technologists. We had two metrics in our survey to evaluate technical background: prior technical event attendance and occupation. First, we found that 71 out of 179 (40%) of participants had never attended a previous event, panel, or talk on data or technology, suggesting that this is their first attendance to an event on a technology-related subject. To complement this high number of non-technologists, we found that of the 151 respondents that specified their career or field of study, 67 (44%) were in computer science or another STEM field. This meant over half of our audience were non-technologists with STEM background at all.

From our demographic questions, we found that there was a significant increase in female participation compared to the 26% of women involved in technical careers [32]. For gender, 89 survey participants identified as female, 85 as male, and 1 as other. Our other demographic metrics showed low ethnic and educational diversity. We found that the majority of the participants were under 30 (66%), college educated (93%), and white (79%). This is mostly an artifact of the place where we ran this performance, which is mostly white and educated.

Another indicator of successful engagement is what a person may feel or want to do following the show. 45.8% of the audience members reported feeling "concerned" after the show. As reported

above, the 62% who were motivated to learn following the show also signals an opportunity for continued engagement in the topics.

We asked audience members about their interest in certain specific activities following the show. As a whole, the audience strongly agreed that they "want to learn more about how companies use and share personal data" ($4.173/5, \sigma = 0.65$). they further agreed that they would "want to attend a panel, talk, or event on data privacy" ($3.775/5, \sigma = 0.88$), and "want to use tools to visualize their own data" ($3.98/5, \sigma = 0.81$). This further indicates that a performance like Quantified Self may be a strong initial engagement where further programming is planned.

We also asked what users would change to improve engagement in a future run of the show. Many people wanted forms of greater interaction (41/103), including more discussion with the actors, increasing data usage by the actors and exhibits, and giving the audience more agency to change the outcomes of the story. Example respondent quotes in this vein include: "**Display audience members info in front of whole audience and let that audience members reaction influence trajectory of show.**"; "**I thought the people/actors would "know" about us, not just the machines/games**".

Others wanted more clarity on how they were supposed to interact (15%) or deeper background information (15%) coming into or during the show : "**Made it clear how much I was supposed to affect the show. For instance, Amelia's reminder on the bookcase. I was unsure if I would unknowingly throw a wrench in the show by interacting with it.**" "**...it would have been cool if there were little clues all around giving deeper, technical, or background info to help us interact more intentionally.**"

5.3.6.5 Listening to our Audience's Attitudes and Beliefs

Allowing our audience to communicate their perspectives to the production team was a further goal. We designed the show given certain ethical thinking and felt it appropriate to listen to the other side. We compiled a list of seven questionable practices by companies and asked them whether they agreed (5) or disagreed (1). This was to validate whether our content was in line with

their concerns.

We found that audience members were most troubled by companies withholding information on how their data is used and least concerned about companies using their data to research new products. Asked whether they "believed it was okay for companies to withhold information about how their data is used" the average response was 1.27 (strongly disagree, $\sigma = 0.61$). A close second was companies withholding information on who data is shared with $(1.31, \sigma = 0.68)$. Surprising to us was that these two withholding practices were rated as more disagreeable than a hacker accessing personal data for a social cause $(2.13, \sigma = 1.086)$ and companies tracking them through their friends' social media accounts $(1.70, \sigma = 0.90)$. The implication here is that transparency is a primary factor for users regarding data ethics.

On average our audience agreed with companies using their data to do research on new products $(3.14/5, \sigma = 1.02)$. The practices carried the most variance among audience members were targeted advertisement using one's personal data (highest, $\sigma = 1.089$) and hackers accessing personal data for social causes (2nd highest, $\sigma = 1.086$), which suggests there exist diverse perspectives.

Digging a bit further, we found some of these concerns differed significantly by whether or not someone had a background in STEM. There was a mean difference of .55 such those with STEM backgrounds more strongly agreed with targeted advertisements $(t = 3.217, p = .002)$. Similarly, we found people with STEM backgrounds more accepting of companies selling their data to third parties (+.37 mean difference, $t = -2.224, p = .028$) and companies using their data to research products (+.4 mean difference, $t = -2.499, p = .014$).

5.3.6.6 Reaction to Exhibits

Our survey had one dedicated question probing the audience sentiments of the technology exhibits. We asked for respondents to list their favorite piece, least favorite piece, and rationale for that choice. As this question was open-ended with a short response, we received less consistent answers. Out of the 179 surveyed audience members, 76 answered with both favorite and least

favorite and 72 answered partially (we assumed that singular was a favorite response). 68 gave a rationale for their positive choice and 62 for their least favorite. Overall, we saw that the preferred exhibits elicited reflective experiences, increased social opportunities, and provided accurate feedback about the user. Disappointment with exhibits predominantly called for further interactive features and clearer technological explanations for learning.

Some of the stated rationale revealed the exhibits triggering reflective experiences for some audience members: "**the own-up table because it forced me to consider whether i would share publicly things that id already shared publicly.**"; "**Favorite: highly recommended. It was interested to visualize how specific data is translated into a more general image of myself**"; "**Favorite - own up. Entertaining, felt alienated to my own words when out of context.**"; "**Find your fantasy was the best because I wanted to play but the user agreements were crazy! It really made me question how far I am willing to share my data.**" Out of the 68 survey takers who gave a rationale for their exhibit preference, 18 brought up a self-reflective moment, 12 new thoughts/knowledge, and 4 social revelations.

How people reacted to discomfort varied. An interesting finding was that some of the audience was attracted to the exhibits that made them feel most uncomfortable: "**favorite: the libido booth; it was uncomfortable and disturbing and made me feel vulnerable.**" The exhibit, You and Your Libido, which was the most provocative (and graphic) had high values for favorite (17/148) and least favorite (21/148). Two others listed ***own up*** as their favorite stating they enjoyed how "uncomfortable" it was.

Common themes of frustration were wanting even more data integration and personalization or not understanding the intent of the exhibit for why an exhibit was a person's least favorite: "**job candidacy? (didn't understand the mechanism of pros and cons)**" "**Least: Mirror - didn't use much of my info**" "**I didn't enjoy the reading room because the content didn't seem very personalized to the participant.**"

An exciting result was the many people most interested in the conversations and social dynamics. Own up, the exhibit with the most social interaction, was by far the most popular

with 59/148 (40%) listing it as their favorite. 20% of survey takers listed social dynamics as their rational for choosing a favorite exhibit. For example: "**Own up was my favorite - it generated live interaction with strangers**". Another person made a general statement that "**[playing] games with others could learn a little about whats important to them**".

5.3.6.7 Engendered Trust

The collection and use of social media data created a challenge in terms of attendee trust. Prior to the show, we had specifically designed a data collection policy and language clearly displayed on our website to reassure users about the use. In our survey, we evaluated this by asking if our production or Google was trusted more with their data and why.

Out of 179 surveys, 150 answered this question with a clear choice. 63% (95/150) said our production, 25% said Google, 3% said neither, and 9% said both.

A common response for those that trusted us was that we were transparent about our intentions. Out of those who trusted our production, 16 specifically mentioned the policy as engendering that trust, while many more mentioned key components (25 brought up the promise to delete data, 10 said not selling data, 7 commented upon the usage). "**"You stated you would delete everything after the show so I trusted you more. Your terms and conditions were easy to read and understand."**" Those who trusted Google cited its high level of professionalism or the fact that Google already has so much information–"**"it already knows everything so what's to**" hide?"

5.3.6.8 Production Crew

5.3.6.9 Education Goals

We had similar education goals for the production crew. We found most crew members gained insights through the production. When asked whether the experiences had made them more knowledgeable about the issues, most (13/15) were affirmative. The majority felt strongly so (8/15). "**Yeah. Absolutely! (P5)**". There is evidence that the experience helped students dispel certain previously held misconceptions. "**I didn't think [companies offering free services] would**

use my data for malicious intent. (P5)" "**It opened my eyes to the fact that [my data] is being used for things, and that it's being bought and sold, and I had no idea that that was a thing before. (P9)**" This suggests that students did gain new knowledge.

However, there were limitations to the knowledge gained. A number of members in the art crew made specific statements that they became more knowledgeable only about the issues of data use but not about the underlying technical skills such as coding (P4) or data analysis (P2), even though that was of interest to them. Certain misconceptions were still held by some crew members. "[**Texting lets you**] **communicate with people easily without having to put all your information out there (P8)**", when in fact texting could still expose one's personal information. Two people (2/15) reported gaining no or minimal new knowledge. One explained that "**I already knew a lot (P14)**" prior to joining the crew. The other explained that "**because I worked more on the front-end of the project, I wasn't doing a lot of the research tasks. (P11)**" But this same subject also reported that "[**the show**] **made me perhaps a little bit more skeptical (P11)**". This seems to suggest while a person might not gain concrete knowledge, the person nevertheless had a change of attitude.

5.3.6.10 Engagement

Several crew members expressed being more keen and equipped to have conversations about data sharing issues after the QSelf production experience. "**I had discussions with my friends and my roommates, ever since the show, about these topics and I think that they shared very similar views to what I had before the show. (P5)**" Moreover, the new gained knowledge allows this crew member to bring the discussions to a deeper level. "**And now afterwards, we can have realistic discussions about big data being a commodity. And how that shapes our society and how that shapes our perception and our interactions with everyone else in it. (P5)**"; "**It's something that I bring up more in conversations... people will randomly make a comment, like an advertisement being eerily close, and I can plug in and talk about Quantified Self, and people are always kind of surprised about how much they're sharing, without even realizing**

it. (P2)" This suggests students have become teachers.

There was an appreciation among many interviewees of how the Quantified Self's balanced focus on art and technology provided them with valuable opportunity to work with the other side. There was evidence of mutual learning between groups. "**What came out of doing the project was knowing more about how tech people or computer scientists actually go about doing this. (P14)**" Learning also occurred during team meetings. "**It came up in one of our meetings, is that I only ever had to sign away my data rights once, for that to apply to all of the companies. (P5)**" Several members commented on the edifying experiences of playing a character interacting with the audience. "**[I] become more a little more articulate, especially 'cause I had to talk to the audience about it...and try to draw, elicit responses from other people so that makes you kind of think through it a little more. (P6)**"

Cast members found it especially engaging to be able to interact with audience members while they stay in characters representing different sides, as exemplified by: "**I was playing a game with [audience members] and they said, how'd you get this information?" I'm like, Oh, you said that we could have it. I was in character. (P10)**"; "**While companies do [manipulating people] on a grander scale through the internet, I was this character who would go around and my job was to convince people that my side was the correct side. (P8)**"; "**The interesting thing for me was seeing how different audience members reacted to [the opinions of my character]. (P6)**"

5.3.7 Discussion

5.3.7.1 Where We Landed

The final production was an enriching and successful experienced both in terms of process and learned results. However, we do not believe we achieved our ideal design. Using our heuristic from above, we plotted our estimation of the final production (represented in Figure 5.6.2). Here you see a number of perceived asymmetries in our execution.

To begin with, we believe the show was biased toward art rather than technology. The experience was more structured by the plot and aesthetics than the technical content. Though there were technological artifacts, artistic representations overshadowed technical ones. The on-set exhibits were often more "social commentary" than showcasing or explaining the capabilities of data-driven technologies. Where we were unable to implement highly-technical systems, whether due to lack of time or team expertise, we were forced to use either simplified algorithms or entertaining content. Thus, the overall presentation of the technology more under the conceit of art rather than representing state-of-the art technical capability.

We also felt the content was more fixed than improvisational. The narrative itself was not malleable in any way besides having topical discussion with the actors and a few peppered in pieces of dialogue that would be generated in the hour before the performance. The exhibits too, were less personalized and more structured. Again, we only had time to put sophisticated techniques behind 2 of the exhibits. This means data mining for most exhibits, for instance, used keywords rather than sophisticated NLP or auto-grouping rather than a clustering analysis. Or the personalization of an exhibit was minor—replaced names or a couple messages off of your feeds.

5.3.7.2 What Worked

There is no doubt that many of the affordances discussed were achieved by this show. Beyond the fun had by all parties, we saw the artistic engagement function as an attractor for people with different backgrounds. The balance of men and women interested in and participating in the project was high. Similarly, we saw the performance house a good balance between people with STEM and non-STEM backgrounds. The curiosity around technology mixed with the entertainment of a theater piece was enough to stir up interest quickly around our campus and city.

The amount of candid, interesting discussion that occurred between people was also encouraging. There was always a significant crowd of people in attendance that would really ask questions and often stimulate other members of the audience. Not just technologists, these active players fell across the board. Sometimes privacy enthusiasts, tech hobbyists, gregarious personalities, dis-

rupters, or lovers of role play, these people would want to be *in the world* and instigating legitimate discussion. After the show each night, many were walking away impacted and excited, as reflected in our survey results.

Perhaps the most promising aspects of this was the reporting of how many people wanted more. The results suggest that surrounding programming including panels, educational talks, and workshops would all gain better traction if paired with such a production.

5.3.7.3 What Could be Improved

There are many things we would change in a re-run of the production and lessons that others should incorporate into their own work. Even during the production process, we recognized trade-offs and compromises being made to meet our timelines and requirements. The overarching experience could have been more detailed and coherent with diegetic engagements before and after the performance. Sending a more messages or information from the characters, being able to play with your data outside of the show, and setting the background of the company and world more could have improved comprehension and participation.

An obvious takeaway for better engagement is to let audience adapt the narrative. Using a more complex interaction design, as production companies like Blast Theory coordinate, could allow for a deep exploration of one's opinions and desired outcomes with respect to relevant and important technical issues. Planning malleable character arcs and allowing the audience to affect them through their interactions with the technology and actors would take the performance to the next level. In our next run, we also want to make the technological experience more coherent and robust. We would like to see the data experience to mirror an in-show social media platform. Audience being able to see their data and discuss with one another through digital means is likely to amplify learning and engagement. This kind of interaction with live data is already trending in artistic circles, and HCI researchers could extend this trend into fascinating scenario explorations and open up possibilities to study social dynamics of networked systems.

In terms of research, had we not been on fast-paced production timeline, we would have

started the IRB process much earlier. While we do not recommend other researchers invade on the audience engagement to insert more instruments, we see benefits in adding a) anonymous behavioral statistics from the exhibits and b) a small pre/post focus group to study more in depth. Both of these were missed potentials for research. For other fields, the post-survey and talk backs could be great sites for learning political views, policy positions, attitudes about technology, and gauging user experiences.

5.3.8 Conclusion

In this paper, we explored a method to engage, educate, and bring together technical and non-technical participants into data ethics discussions. Aiming for a balance between technical and artistic, and fixed and improvised modes of interaction, we developed Quantified Self: Immersive Theater and Data Experience. We found that mixing immersive theater with interactive, social-media-driven technology exhibits created opportunities for multiple forms of engagement, although many users wanted even greater interaction and personalization. Drawing from this work, we hope to see future research incorporate elements to engage non-technologists in ethical technology design and discussions.

Chapter 6

Conclusion

6.1 Reflections on the Work

6.1.1 The Growing Complexity of the Problem

As the final pages of this work are typed, US society is under a grueling turmoil over the use of social media by Russian agents to influence the 2016 election. Activists who were tricked by fake accounts to engage in protest and vitriol are now saying, "Facebook should take responsibility. Don't find out after the fact. After the fact is too late" [205]. The piece this quote comes from also highlights and questions the fact that Facebook can identify and bring down piracy videos within minutes, but appears to lack capacity or interest in identifying and stopping fake accounts and information. Only days prior to this piece, the writer of the New York Times "State of the Art" column about new technology, a person doing technology journalism for 20 years, openly stated that new technology now comes "freighted with worry." [129]. The writer's claim is that this past year proves that we have not grappled with the amount of control large tech companies have over our lives and that we are only just starting to understand the terrible sides of technologies.

None of this should come as a big surprise. Especially if you read the chapters leading up to this conclusion. Through all technology's dazzle and excitement, it is evolving more rapidly than human institutions, cultures, and perhaps minds, are equipped to control. At this very moment, companies like Uber and Google are working to automate driving. One impact of this would be a great improvement in the efficiency of shipping and trucking industries. Meanwhile, as of October

20, 2017, the Bureau of Labor Statistics tracks nearly 1.5 million jobs in the trucking industry [12]. One might ask: What exactly is the plan here? Are we really prepared to automate away even half of the employees in this massive industry? And this is not even one of our complex technical-ethical challenges such as uninterpretable machine learning models that house bias. This is a straightforward question of, What is the trajectory of this technology and how will it impact our society? Who will win? Who will lose? Are we destining ourselves for another explosive disruption where people are angrily asking, "Why didn't we think ahead?"

What I'd now like to suggest is that this problem of thinking ahead of the pace of technology's imposed changes *is* the problem of our time. Whether it's climate change, unemployment, or a massive data leak that puts everyone's financial security and identities at risk (ie, the recent Equifax leak), we are dealing with a less-than-stable world where our problems are emerging out of vast uncertainty and impetuous decision-making. This is nothing new given historical events like The Irish Potato Famine, The Bubonic Plague, or The Dust Bowl which all took massive tolls on earlier societies due to forces we could not comprehend, predict, or control. However, the great irony now, as policy and risk researcher Sheila Jasanoff claims [174], is that our risks are happening directly in the face of our modern scientific institutions that were meant to clarify and mitigate the problems of yesteryear. In fact, as she argues, and this thesis confers in light of the ethical problems raised by Big Data and Machine Intelligence, our modern scientific and technical institutions are now creating the risks we all take on. The even more upsetting trend, as pointed out by risk theorist Ulrich Beck, is that we are also now being forced to recede our ideals (e.g., privacy, civil liberties) in order to at the same time protect them [47]. Consider here the argument that we cannot over commit to privacy protocols because then terrorists would not be trackable.

Thus, with no more words of worry to harrow our more hopeful aim, it is my urgent suggestion we begin entertaining ideas, criticisms, and projects that are beyond the normal confines of research and market demands. As this thesis has pointed out, one of these possible modes of inquiry and discourse utilizes narrative to aid in negotiating the future of our society. Particularly as it pertains to technology.

6.1.2 The Promises of Narrative and What's Comes Next

This thesis forks off of a deep and broad history of narrative in our cultural and intellectual traditions. Even the most novel of forms of narrative I employed, the immersive theater piece Quantified Self, was pre-empted by contemporary art and theater companies such as Blast Theory, Third Rail, and Punchdrunk Productions. Each of whom have explored the form and style of immersive and interactive theatrical environments. The primary punch of this thesis was to work through a number of domains struggling with the ethical raised by Big Data and Machine Intelligence by applying narrative as a multi-faceted strategy.

We saw narrative as a useful tool for grounding concerns of machines violating human autonomy in Chapter 3. What should be noted from that piece is that a robust discussion of autonomy and machine intervention may have had to stop at the privacy and discrimination lenses, since those are the ones we best understand. Given that we still do not understand the broader impacts of machine intervention, such as those we are picking up the pieces of after our 2016 US election, it was only through fictions that we can see the real scope of the problem. It seems likely that if we had to wait until machine actions completely instrumentalized human actions, it would potentially be too late (or much harder) to go back and fight for our lost autonomy. Chapters 4 and 5 further showed that working with robust socio-technical narratives was an eye opening experience for people of varied backgrounds. Importantly, as was learned in the interviews with Quantified Self cast and crew, many people already were aware of some of the facts around invasion of privacy and targeted advertising. However, it was not until they really wrestled with these facts within a human story and a set of tangible consequences that some were able to really form an opinion or know how they wanted to react.

Of course it is not all so simple. Conceptualizing narratives that are at the same time realistic and poignant takes time, effort, and talent. If we consider the case of the movie "WarGames," we see a way in which narrative can give rise to moral panic and ill-formed protections. The 1983 movie tells a story of young man who is nearly causes a war by accidentally meddling with computers at

NORAD using the nascent internet. This movie startled then-president Ronald Reagan and began a process toward regulating cybersecurity [181]. Though the lasting consequences of overly-reactive policy have been felt in cases like Aaron Schwartz and Kevin Mittnick where the fear of computer insecurity has led to harsh prison terms and forced solitary confinement due to misunderstandings and treating computer crimes as more severe than even violent crimes [107].

For this reason, I think the primary objective of continued research should be to dig into intentional narrative building with multi-disciplinary teams equipped to tease out details and test ideas against reality. The next level of exploration, building from projects like Quantified Self, is toward well-rounded research programs that look at public attitudes, create spaces for debate and disagreement, and offer experiences online and in real life to explore the depths. Treating narrative as a center for such projects allows the ideas to be communicated across communities and develop a central theme to refer back to and ground further inquiry. Projects like this will be critical to extending our conversations about the future outside of expert circles and academic enclaves.

It would be exciting to see policy and legal researchers working alongside sociologists, writers, and engineers to try and formulate preventative mechanisms for future harms. The narratives again acting as a translation tool where boundary scenarios could be drafted, argued for/against, and tweaked to try and discern what limitations might be required to protect our society from developments we do not fully understand. Similarly within organization such as Facebook or Google, it could be imagined that ethics councils may work aside product and feature development teams to determine if new projects may have negative consequences. Once again we may leverage narrative to help support a rubric for what we are trying to avoid or accomplish and hold up these standards to the technology prior to release.

Perhaps most important to developing this research further is to sort out how narrative can best support a site of negotiation. That is, how can we productively use narrative to allow diverse stakeholders to test out and discuss what they are comfortable and uncomfortable with happening? Is it possible through performance, group interviews, or some other means to devise situations where we can meaningfully present alternative futures and make sense of what possibilities we are trying

to fend off or go toward. It seems to me that experts, broadly, are in need of finding a way to take stances against unacceptable or worrisome developments in their field. Often this is difficult due to the apolitical stance research strives to hold. However, it seems narrative-based research may help us determine the future charter where the public, or a particular group's, interests are most at risk or the most benefited and use these researched narratives to support taking a stand when we reach questionable junctures. This is the very argument made by author and futurist David Brin in his essay "The Self-Preventing Prophecy" [67]. In it he claims that 1984 did important work to prevent us from quickly jumping into a pure surveillance state as soon as it was technologically feasible due to the vision of Big Brother and our collective desire to avoid this fate of humanity. Whether or not he is right, it seems almost too obvious that we all can meaningfully say that depictions of SkyNet, Big Brother, or the many disturbing technological-tales in Black Mirror are all destinations we hope to avoid. The questions are now, "How do we know we are headed that direction?" and "How do we prevent it?"

Though this dissertation is only a few of the first steps toward formalizing a relationship between ethical inquiry and narrative within a research program, the possibilities are endless. In the coming years it is my hope to continue formalizing these methods and work deeper in the community setting to understand how best to create constructive multi-stakeholder inquiry. Until then, I hope this thesis project and these words are able to impel researchers to consider using narrative to help them see beyond their imagined best-cases and noble stakeholders who will use their technologies with the highest moral regard. Otherwise it may, in hindsight, appear chillingly prescient that our current culture obsessed by genetically-mutated zombies and apocalyptic circumstances was correct in that the only story it knew, the only option it saw, was one of disaster. Rather than obsess on the fantastical, let's instead work to discover and pursue the story of a society that takes the reins of technology into the hands of many to benefit the most possible.

6.2 Reader's Guide

Ending this thesis, it strikes me that it's unlikely a casual reader interested in narrative and ethics who comes across this work will actually read this document cover-to-cover. So here as an ending remark, I am offering a brief reader's guide that may help someone selectively read sections relevant to their pursuits. My best advice would be for all readers to sink the time into the first chapter, but from there:

- **HCI Researchers**: Chapter 4 (particularly sections 1 and 4) and Chapter 5 is most pertinent to seeing narrative applied to problems and related to methods most commonly studied within HCI.
- **Policy and Legal Scholars**: Chapters 2 and 3 both provide context for policy and law as well as show how narrative can be used to build policy frameworks for technology issues.
- **Artists and Writers**: Chapters 4 (section 1) and Chapter 5 (section 2) are examples of how creative writing and artistic storytelling can be useful for intentionally engaging the public imagination on these matters.
- **Educators**: Chapter 4 (sections 2 and 3) are direct discussions of the educational impact narrative can have within classrooms and on multi-disciplinary projects.
- **Enthusiast**: Chapter 2 is where you can learn the background of these ethical problems and Chapter 5 may provide the most insight into how the public is thinking about these matters.

Bibliography

[1] Federal Trade Commission v. Vizio Inc.

[2] Predict Crime | Predictive Policing Software.

[3] Self-regulatory principles for online behavioral advertising. Technical report, Federal Trade Commission, February 2009.

[4] Facebook 'likes' predict personality. BBC News, March 2013.

[5] Google apologises for Photos app's racist blunder. BBC News, July 2015.

[6] Overview of the Privacy Act of 1974. Technical report, United States Department of Justice, 2015.

[7] Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues. Technical report, Federal Trade Commission, Washington DC, 2016.

[8] Computer Engineering Curricula 2016. Technical report, Joint Task Force on Computer Engineering Curricula, December 2016.

[9] Criteria for Accrediting Computing Programs, 2017-2018. Technical report, ABET, Baltimore, MD, 2016.

[10] New DDoS attack technique could unleash devastating internet meltdown warn experts, October 2016.

[11] Google and the Miseducation of Dylann Roof, January 2017.

[12] Industries at a Glance: Truck Transportation: NAICS 484, October 2017.

[13] VIZIO to Pay $2.2 Million to FTC, State of New Jersey to Settle Charges It Collected Viewing Histories on 11 Million Smart Televisions without Users Consent, February 2017.

[14] Alessandro Acquisti. Price Discrimination, Privacy Technologies, and User Acceptance. In CHI Workshop on Personalization and Privacy. Citeseer, 2006.

[15] Alessandro Acquisti and Ralph Gross. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In International workshop on privacy enhancing technologies, pages 36–58. Springer, 2006.

[16] Eytan Adar, Carolyn Gearig, Ayshwarya Balasubramanian, and Jessica Hullman. PersaLog: Personalization of News Article Content. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, pages 3188–3200, New York, NY, USA, 2017. ACM.

[17] Julius Adebayo and Lalana Kagal. Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box Models. arXiv preprint arXiv:1611.04967, 2016.

[18] Ameeta Agrawal and Aijun An. Unsupervised Emotion Detection from Text Using Semantic and Syntactic Relations. In Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01, WI-IAT '12, pages 346–353, Washington, DC, USA, 2012. IEEE Computer Society.

[19] Philip E Agre. Surveillance and Capture: Two Models of Privacy. The Information Society, 10.2:101–127, 1994.

[20] Aws Albarghouthi, Loris D'Antoni, Samuel Drews, and Aditya Nori. Fairness as a Program Property. arXiv:1610.06067 [cs], October 2016. arXiv: 1610.06067.

[21] Guy-Alain Amoussou, Myles Boylan, and Joan Peckham. Interdisciplinary Computing Education for the Challenges of the Future. In Proceedings of the 41st ACM Technical Symposium on Computer Science Education, SIGCSE '10, pages 556–557, New York, NY, USA, 2010. ACM.

[22] M. Ananny. Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. Science, Technology & Human Values, 41(1):93–117, January 2016.

[23] Josue Anaya and Adrian Barbu. RENOIR - A Dataset for Real Low-Light Image Noise Reduction. arXiv:1409.8230 [cs], September 2014. arXiv: 1409.8230.

[24] Ronald E Anderson. ACM code of ethics and professional conduct. Communications of the ACM, 35(5):94–99, 1992.

[25] Mark Andrejevic. Big Data, Big Questions| The Big Data Divide. International Journal of Communication, 8(0):17, June 2014.

[26] Julio Angulo and Martin Ortlieb. WTH..! Experiences, reactions, and expectations related to online privacy panic situations. In Symposium on Usable Privacy and Security (SOUPS), 2015.

[27] Julia Angwin, Surya Mattu, Jeff Larson, and Lauren Kirchner. Machine Bias: Theres Software Used Across the Country to Predict Future Criminals. And its Biased Against Blacks., May 2016.

[28] Dunne Anthony and Fiona Raby. Speculative Everything: Design, Fiction, and Social Dreaming. London, England: The Mit Press, Cambridge, Massachusetts, 2013.

[29] Annie I. Antn, Julia B. Earp, and Jessica D. Young. How internet users' privacy concerns have evolved since 2002. IEEE Security & Privacy, 8(1), 2010.

[30] Anne G. Applin. A Learner-centered Approach to Teaching Ethics in Computing. In Proceedings of the 37th SIGCSE Technical Symposium on Computer Science Education, SIGCSE '06, pages 530–534, New York, NY, USA, 2006. ACM.

[31] Christie Aschwanden and Ritchie King. Science Isnt Broken, August 2015.

[32] Catherine Ashcraft, Brad McLain, and Elizabeth Eger. Women in Tech: The Facts (2015-16 Update). Technical report, National Center for Women and Information Technology.

[33] James Auger. Speculative design: crafting the speculation. Digital Creativity, 24(1):11–35, March 2013.

[34] Yannis Bakos, Florencia Marotta-Wurgler, and David Trossen. Does Anyone Read the Fine Print? Consumer Attention to Standard Form Contracts. New York University Law and Economics Working Papers, January 2014.

[35] Rebecca Balebako, Jaeyeon Jung, Wei Lu, Lorrie Faith Cranor, and Carolyn Nguyen. "Little Brothers Watching You": Raising Awareness of Data Leaks on Smartphones. In Proceedings of the Ninth Symposium on Usable Privacy and Security, SOUPS '13, pages 12:1–12:11, New York, NY, USA, 2013. ACM.

[36] Martina Balestra, Orit Shaer, Johanna Okerlund, Madeleine Ball, and Oded Nov. The Effect of Exposure to Social Annotation on Online Informed Consent Beliefs and Behavior. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW '16, pages 900–912, New York, NY, USA, 2016. ACM.

[37] Michael Barbaro and Tom Zeller Jr. A Face Is Exposed for AOL Searcher No. 4417749. The New York Times, August 2006.

[38] Solon Barocas. The price of precision: Voter microtargeting and its potential harms to the democratic process. In Proceedings of the first edition workshop on Politics, elections and data, pages 31–36. ACM, 2012.

[39] Solon Barocas. Data Mining and the Discourse on Discrimination. In Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining, 2014.

[40] Solon Barocas and Helen Nissenbaum. On notice: The trouble with Notice and Consent. In Proceedings of the Engaging Data Forum: The First International Forum on the Application and Management of Personal Electronic Information, 2009.

[41] Solon Barocas and Helen Nissenbaum. Big datas end run around anonymity and consent. Privacy, big data, and the public good: Frameworks for Engagement, pages 44–75, 2014.

[42] Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. SSRN Scholarly Paper ID 2477899, Social Science Research Network, Rochester, NY, August 2014.

[43] Anna Maria Barry-Jester. Should Prison Sentences Be Based On Crimes That Havent Been Committed Yet?, August 2015.

[44] Rebecca Bates, Judy Goldsmith, Rosalyn Berne, Valerie Summet, and Nanette Veilleux. Science Fiction in Computer Science Education. In Proceedings of the 43rd ACM Technical Symposium on Computer Science Education, SIGCSE '12, pages 161–162, New York, NY, USA, 2012. ACM.

[45] Eric PS Baumer, June Ahn, Mei Bie, Elizabeth M. Bonsignore, Ahmet Brtecene, Ouz Turan Buruk, Tamara Clegg, Allison Druin, Florian Echtler, and Dan Gruen. CHI 2039: speculative research visions. In CHI'14 Extended Abstracts on Human Factors in Computing Systems, pages 761–770. ACM, 2014.

[46] Tom L. Beauchamp and James F. Childress. Principles of Biomedical Ethics. Oxford University Press, Oxford, 1979.

[47] Ulrich Beck. Living in the world risk society: A Hobhouse Memorial Public Lecture given on Wednesday 15 February 2006 at the London School of Economics. Economy and Society, 35(3):329–345, August 2006.

[48] Kim Bellware. Uber Settles Investigation Into Creepy 'God View' Tracking Program. Huffington Post, January 2016.

[49] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, and Aleix M. Martinez. EmotioNet Challenge: Recognition of facial expressions of emotion in the wild. arXiv:1703.01210 [cs], March 2017. arXiv: 1703.01210.

[50] Stanley I. Benn. Privacy, freedom, and respect for persons. In Ferdinand David Schoeman, editor, Philosophical Dimensions of Privacy: An Anthology, pages 223–244. Cambridge University Press, Cambridge, 1984.

[51] Stanley I. Benn. A Theory of Freedom. Cambridge University Press, Cambridge, 1988.

[52] Andrew Besmer and Heather Richter Lipford. Users'(mis) conceptions of social applications. In Proceedings of Graphics Interface 2010, pages 63–70. Canadian Information Processing Society, 2010.

[53] Binod Bhattarai, Gaurav Sharma, and Frdric Jurie. CP-mtML: Coupled projection multi-task metric learning for large scale face retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4226–4235, 2016.

[54] Johana Bhuiyan and Charlie Warzel. "God View": Uber Investigates Its Top New York Executive For Privacy Violations, November 2014.

[55] Igor Bilogrevic and Martin Ortlieb. "If You Put All The Pieces Together...": Attitudes Towards Data Combination and Sharing Across Services and Companies. pages 5215–5227. ACM Press, 2016.

[56] Colin Bird. Status, Identity, and Respect. Political THeory, 32(2):207–232, April 2004.

[57] Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach. Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI. SSRN Scholarly Paper ID 2846909, Social Science Research Network, Rochester, NY, October 2016.

[58] Julian Bleecker. Design Fiction: A short essay on design, science, fact and fiction. Near Future Laboratory, 29, 2009.

[59] Everett Franklin Bleiler. Science-fiction, the early years: a full description of more than 3,000 science-fiction stories from earliest times to the appearance of the genre magazines in 1930: with author, title, and motif indexes. Kent State University Press, 1990.

[60] Mark Blythe. The Hitchhiker's Guide to Ubicomp: Using Techniques from Literary and Critical Theory to Reframe Scientific Agendas. Personal Ubiquitous Comput., 18(4):795–808, April 2014.

[61] Mark Blythe. Research Through Design Fiction: Narrative in Real and Imaginary Abstracts. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14, pages 703–712, New York, NY, USA, 2014. ACM.

[62] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. arXiv:1607.06520 [cs, stat], July 2016. arXiv: 1607.06520.

[63] Richard J. Botting. Teaching and Learning Ethics in Computer Science: Walking the Walk. In Proceedings of the 36th SIGCSE Technical Symposium on Computer Science Education, SIGCSE '05, pages 342–346, New York, NY, USA, 2005. ACM.

[64] Kevin W. Bowyer and Patrick J. Flynn. The ND-IRIS-0405 Iris Image Dataset. arXiv:1606.04853 [cs], June 2016. arXiv: 1606.04853.

[65] danah boyd and Kate Crawford. CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5):662–679, June 2012.

[66] Ben Brantley. Sleep No More Is a Macbeth in a Hotel - Review. The New York Times, April 2011.

[67] David Brin. The self-preventing prophecy. Or, how a dose of nightmare can help tame tomorrows perils. On nineteen eighty-four. Orwell and our future, pages 222–230, 2005.

[68] Bo Brinkman and Keith W. Miller. The Code of Ethics Quiz Show. In Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, SIGCSE '17, pages 679–680, New York, NY, USA, 2017. ACM.

[69] Frederick P. Brooks, Jr. The Teacher's Job is to Design Learning Experiences; Not Primarily to Impart Information. In Proceedings of the 43rd ACM Technical Symposium on Computer Science Education, SIGCSE '12, pages 1–2, New York, NY, USA, 2012. ACM.

[70] Simone Browne. Dark matters: On the surveillance of blackness. Duke University Press, 2015.

[71] Erik Brynjolfsson and Andrew McAfee. The second machine age: Work, progress, and prosperity in a time of brilliant technologies. WW Norton & Company, 2014.

[72] Joy Buolamwini. AJL -ALGORITHMIC JUSTICE LEAGUE.

[73] Joy Buolamwini. The Algorithmic Justice League, December 2016.

[74] Joy Buolamwini. How I'm fighting bias in algorithms, November 2016.

[75] Ken Burns, Dayton Duncan, and Julie Dunfey. The Dust Bowl: A Film by Ken Burns. Pbs. org, 2012.

[76] Sarah Buss. Respect for Persons. Canadian Journal of Philosophy, 29(4):517–550, December 1999.

[77] Rainer Bhme and Stefan Kpsell. Trained to Accept?: A Field Experiment on Consent Dialogs. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10, pages 2403–2406, New York, NY, USA, 2010. ACM.

[78] Mary Elaine Califf and Mary Goodwin. Effective Incorporation of Ethics into Courses That Focus on Programming. In Proceedings of the 36th SIGCSE Technical Symposium on Computer Science Education, SIGCSE '05, pages 347–351, New York, NY, USA, 2005. ACM.

[79] Stuart Candy. The Futures of Everyday Life: Politics and the Design of Experiential Scenarios. PhD thesis, University of Hawaii at Manoa, 2010.

[80] Stuart Candy and Jake Dunagan. Designing an experiential scenario: The People Who Vanished. Futures, June 2016.

[81] John Carroll, Mary Beth Rosson, and Paul McInerney. Scenarios in Practice. In CHI '03 Extended Abstracts on Human Factors in Computing Systems, CHI EA '03, pages 1046–1047, New York, NY, USA, 2003. ACM.

[82] John M. Carroll, editor. Scenario-based Design: Envisioning Work and Technology in System Development. John Wiley & Sons, Inc., New York, NY, USA, 1995.

[83] Lukas Cavigelli, Dominic Bernath, Michele Magno, and Luca Benini. Computationally efficient target classification in multispectral image data with Deep Neural Networks. In SPIE Security+ Defence, pages 99970L–99970L. International Society for Optics and Photonics, 2016.

[84] L. Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K. Vishnoi. How to be Fair and Diverse? arXiv:1610.07183 [cs], October 2016. arXiv: 1610.07183.

[85] Chris Chambers. The psychology of mass government surveillance: How do the public respond and is it changing our behaviour? The Guardian, March 2015.

[86] Stevie Chancellor, Zhiyuan Lin, Erica L. Goodman, Stephanie Zerwas, and Munmun De Choudhury. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW '16, pages 1171–1184, New York, NY, USA, 2016. ACM.

[87] Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. Quantitative Analysis of Automatic Image Cropping Algorithms: A Dataset and Comparative Study. arXiv:1701.01480 [cs], January 2017. arXiv: 1701.01480.

[88] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv preprint arXiv:1703.00056, 2017.

[89] Benedict Chukuka and Michael Locasto. A Survey of Ethical Agreements in Information Security Courses. In Proceedings of the 47th ACM Technical Symposium on Computing Science Education, SIGCSE '16, pages 479–484, New York, NY, USA, 2016. ACM.

[90] Nicholas Confessore and Danny Hakim. Data Firm Says Secret Sauce Aided Trump; Many Scoff. The New York Times, March 2017.

[91] Melissa Cote and Alexandra Branzan Albu. Teaching Computer Vision and Its Societal Effects: A Look at Privacy and Security Issues From the Students' Perspective. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, July 2017.

[92] Tim Coughlan, Michael Brown, Glyn Lawson, Richard Mortier, Robert J. Houghton, and Murray Goulden. Tailored Scenarios: A Low-cost Online Method to Elicit Perceptions on Designs Using Real Relationships. In CHI '13 Extended Abstracts on Human Factors in Computing Systems, CHI EA '13, pages 343–348, New York, NY, USA, 2013. ACM.

[93] National Research Council and others. Understanding risk: Informing decisions in a democratic society. National Academies Press, 1996.

[94] K. Crawford. Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics. Science, Technology & Human Values, 41(1):77–92, January 2016.

[95] Kate Crawford. The Hidden Biases in Big Data, April 2013.

[96] Kate Crawford and Jason Schultz. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. SSRN Scholarly Paper ID 2325784, Social Science Research Network, Rochester, NY, October 2013.

[97] Kamal Dahbur, Bassil Mohammad, and Ahmad Bisher Tarakji. A Survey of Risks, Threats and Vulnerabilities in Cloud Computing. In Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications, ISWSA '11, pages 12:1–12:6, New York, NY, USA, 2011. ACM.

[98] Nicholas S. Dalton, Rebecca Moreau, and Ross K. Adams. Resistance is Fertile: Design Fictions in Dystopian Worlds. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA '16, pages 365–374, New York, NY, USA, 2016. ACM.

[99] Stephen Darwall. The Second-Person Standpoint. Harvard University Press, Cambridge, MA, 2006.

[100] Stephen L. Darwall. Two Kinds of Respect. Ethics, 88(1):36–49, October 1977.

[101] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated Experiments on Ad Privacy Settings. Proceedings on Privacy Enhancing Technologies, 2015(1):92–112, 2015.

[102] Thomas H. Davenport and D. J. Patil. Data Scientist: The Sexiest Job of the 21st Century, October 2012.

[103] David J. Hauser and Norbert Schwarz. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. Behavior Research Methods, 48(1):400–407, March 2016.

[104] Janet Davis and Henry M. Walker. Incorporating Social Issues of Computing in a Small, Liberal Arts College: A Case Study. In Proceedings of the 42Nd ACM Technical Symposium on Computer Science Education, SIGCSE '11, pages 69–74, New York, NY, USA, 2011. ACM.

[105] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Predicting Postpartum Changes in Emotion and Behavior via Social Media. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, pages 3267–3276, New York, NY, USA, 2013. ACM.

[106] Benedict de Spinoza. Ethics. Simon and Schuster, 1970.

[107] Declan McCullagh. From 'WarGames' to Aaron Swartz: How U.S. anti-hacking law went astray, March 2013.

[108] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 248–255. IEEE, 2009.

[109] John Dewey and Melvin L Rogers. The public and its problems: An essay in political inquiry. Penn State Press, 2012.

[110] Scott Dexter, Elizabeth Buchanan, Kellen Dins, Kenneth R. Fleischmann, and Keith Miller. Characterizing the Need for Graduate Ethics Education. In Proceeding of the 44th ACM Technical Symposium on Computer Science Education, SIGCSE '13, pages 153–158, New York, NY, USA, 2013. ACM.

[111] Robin S. Dillon. Respect. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

[112] Carl DiSalvo. Spectacles and Tropes: Speculative Design and Contemporary Food Cultures. The Fibreculture Journal, (20 2012: Networked Utopias and Speculative Futures), 2012.

[113] Alan Donagan. The Theory of Morality. University of Chicago Press, Chicago, 1977.

[114] Paul Dourish. What we talk about when we talk about context. Personal and ubiquitous computing, 8(1):19–30, 2004.

[115] Paul Dourish and Genevieve Bell. Resistance is futile: reading science fiction alongside ubiquitous computing. Personal and Ubiquitous Computing, 18(4):769–778, April 2014.

[116] E. Dowell and E. Weitkamp. An exploration of the collaborative processes of making theatre inspired by science. Public Understanding of Science, 21(7):891–901, October 2012.

[117] R. S. Downie and Elizabeth Telfer. Respect for persons. Schocken Books, New York, 1969.

[118] Anthony Dunne and Fiona Raby. Speculative everything: design, fiction, and social dreaming. MIT Press, 2013.

[119] Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4):211–407, 2013.

[120] Laurel Eckhouse. Big data may be reinforcing racial bias in the criminal justice system. The Washington Post, February 2017.

[121] Serge Egelman. My Profile is My Password, Verify Me!: The Privacy/Convenience Trade-off of Facebook Connect. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, pages 2369–2378, New York, NY, USA, 2013. ACM.

[122] Serge Egelman, Julia Bernd, Gerald Friedland, and Dan Garcia. The Teaching Privacy Curriculum. In Proceedings of the 47th ACM Technical Symposium on Computing Science Education, SIGCSE '16, pages 591–596, New York, NY, USA, 2016. ACM.

[123] Mostafa El-Bermawy. Your Filter Bubble is Destroying Democracy. WIRED, November 2016.

[124] Chris Elsden, David Chatting, Abigail C. Durrant, Andrew Garbett, Bettina Nissen, John Vines, and David S. Kirk. On Speculative Enactments. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, pages 5386–5399, New York, NY, USA, 2017. ACM.

[125] Chris Elsden, Bettina Nissen, Andrew Garbett, David Chatting, David Kirk, and John Vines. Metadating: Exploring the Romance and Future of Personal Data. pages 685–698. ACM Press, 2016.

[126] Richard G. Epstein. An Ethics and Security Course for Students in Computer Science and Information Technology. In Proceedings of the 37th SIGCSE Technical Symposium on Computer Science Education, SIGCSE '06, pages 535–537, New York, NY, USA, 2006. ACM.

[127] Robert Epstein and Ronald E. Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. Proceedings of the National Academy of Sciences, 112(33):E4512–E4521, August 2015.

[128] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5562–5570, 2016.

[129] Farhad Manjoo. Why Tech Is Starting to Make Me Uneasy. The New York Times, October 2017.

[130] Mauricio S. Featherman and John D. Wells. The Intangibility of e-Services: Effects on Perceived Risk and Acceptance. SIGMIS Database, 41(2):110–131, May 2010.

[131] Joel Feinberg. The nature and value of rights. The Journal of Value Inquiry, 4(4):243–260, December 1970.

[132] Miguel Ferreira, Muhammad Bilal Zafar, and Krishna P Gummadi. The Case for Temporal Transparency: Detecting Policy Change Events in Black-Box Decision Making Systems. arXiv preprint arXiv:1610.10064, 2016.

[133] Jacquelyn S. Fetrow and David J. John. Bioinformatics and Computing Curriculum: A New Model for Interdisciplinary Courses. In Proceedings of the 37th SIGCSE Technical Symposium on Computer Science Education, SIGCSE '06, pages 185–189, New York, NY, USA, 2006. ACM.

[134] Casey Fiesler, Michaelanne Dye, Jessica L. Feuston, Chaya Hiruncharoenvate, C.J. Hutto, Shannon Morrison, Parisa Khanipour Roshan, Umashanthi Pavalanathan, Amy S. Bruckman, Munmun De Choudhury, and Eric Gilbert. What (or Who) Is Public?: Privacy Settings and

Social Media Content Sharing. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17, pages 567–580, New York, NY, USA, 2017. ACM.

[135] Casey Fiesler, Cliff Lampe, and Amy S Bruckman. Reality and Perception of Copyright Terms of Service for Online Content Creation. pages 1448–1459. ACM Press, 2016.

[136] Baruch Fischhoff, Paul Slovic, Sarah Lichtenstein, Stephen Read, and Barbara Combs. How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. Policy sciences, 9(2):127–152, 1978.

[137] Baruch Fischhoff, Stephen R. Watson, and Chris Hope. Defining risk. Policy Sciences, 17(2):123–139, 1984.

[138] Philippa Foot. The Problem of Abortion and the Doctrine of the Double Effect. Oxford Review, 5:5–15, 1967.

[139] William K. Frankena. The Ethics of Respect for Persons. Philosophical Topics, 14(2):149–167, 1986.

[140] Harry Frankfurt. Equality and Respect. Social Research, 64(1):3–15, 1997.

[141] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. arXiv:1609.07236 [cs, stat], September 2016. arXiv: 1609.07236.

[142] Alexander R Galloway. Laruelle: Against the digital. University of Minnesota Press Minneapolis, 2014.

[143] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. The Perpetual Line-Up. Technical report, Center on Privacy and Technology, Georgetown Law, 2016.

[144] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles. Automatic Age Estimation Based on Facial Aging Patterns. IEEE Trans. Pattern Anal. Mach. Intell., 29(12):2234–2240, December 2007.

[145] Samuel Gibbs. Musk, Wozniak and Hawking urge ban on warfare AI and autonomous weapons. The Guardian, July 2015.

[146] Samuel Gibbs. Women less likely to be shown ads for high-paid jobs on Google, study shows. The Guardian, July 2015.

[147] Tarleton Gillespie. The Relevance of Algorithms. Media technologies: Essays on communication, materiality, and society, page 167, 2014.

[148] Sarah Glaz and Su Liang. Modelling with poetry in an introductory college algebra course and beyond. Journal of Mathematics and the Arts, 3:123–133, September 2009.

[149] Jennifer Golbeck. All Eyes On You, September 2014.

[150] Debra S. Goldberg and Elizabeth K. White. E Pluribus, Plurima: The Synergy of Interdisciplinary Class Groups. In Proceedings of the 45th ACM Technical Symposium on Computer Science Education, SIGCSE '14, pages 457–462, New York, NY, USA, 2014. ACM.

[151] Bryce Goodman and Seth Flaxman. EU regulations on algorithmic decision-making and a "right to explanation". arXiv:1606.08813 [cs, stat], June 2016. arXiv: 1606.08813.

[152] Jonny Griffiths. Art by numbers: a collaboration between an Art and Maths Department. Journal of Mathematics and the Arts, 3:155–170, September 2009.

[153] Frances Grodzinsky, Ed Gehringer, Laurie S. King, and Herman Tavani. Responding to the Challenges of Teaching Computer Ethics. In Proceedings of the 35th SIGCSE Technical Symposium on Computer Science Education, SIGCSE '04, pages 280–281, New York, NY, USA, 2004. ACM.

[154] Jonathan Grudin. Desituating action: Digital representation of context. Human-Computer Interaction, 16(2):269–286, 2001.

[155] Jurgen Habermas. Between facts and norms: Contributions to a discourse theory of law and democracy. Mit Press, 1996.

[156] Abdenour Hadid. Face biometrics under spoofing attacks: Vulnerabilities, countermeasures, open issues, and research directions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 113–118, 2014.

[157] Brian R. Hall. A Synthesized Definition of Computer Ethics. SIGCAS Comput. Soc., 44(3):21–35, October 2014.

[158] Bernard E Harcourt. Against prediction: Profiling, policing, and punishing in an actuarial age. University of Chicago Press, 2008.

[159] Brian Harrell. Protecting vital electricity infrastructure, March 2016.

[160] Tristan Harris. How Technology is Hijacking Your Mind from a Former Insider, May 2016.

[161] Behzad Hasani and Mohammad H. Mahoor. Spatio-Temporal Facial Expression Recognition Using Convolutional Neural Networks and Conditional Random Fields. arXiv:1703.06995 [cs], March 2017. arXiv: 1703.06995.

[162] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs], December 2015. arXiv: 1512.03385.

[163] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data. arXiv:1511.05284 [cs], November 2015. arXiv: 1511.05284.

[164] Joseph R. Herkert. Engineering ethics education in the USA: Content, pedagogy and curriculum. European Journal of Engineering Education, 25(4):303–313, December 2000.

[165] Tony Hey, Stewart Tansley, Kristin M Tolle, and others. The fourth paradigm: data-intensive scientific discovery, volume 1. Microsoft Research, Redmond, WA, 2009.

[166] D Sunshine Hillygus and Todd G Shields. The persuadable voter: Wedge issues in presidential campaigns. Princeton University Press, 2014.

[167] Paul Hitlin. Research in the Crowdsourcing Age, a Case Study, July 2016.

[168] Jessica Hodgins. Educating for Both Art and Technology. In Proceedings of the 46th ACM Technical Symposium on Computer Science Education, SIGCSE '15, pages 1–1, New York, NY, USA, 2015. ACM.

[169] Eric Holder. Attorney General Eric Holder Speaks at the National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference, August 2014.

[170] Luke Hutton and Tristan Henderson. " I didn't sign up for this!": Informed consent in social network research. In Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM), 2015.

[171] Lucas D. Introna. Maintaining the reversibility of foldings: Making the ethics (politics) of information technology visible. Ethics and Information Technology, 9(1):11–25, February 2007.

[172] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fair Learning in Markovian Environments. arXiv:1611.03071 [cs], November 2016. arXiv: 1611.03071.

[173] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in Reinforcement Learning. arXiv:1611.03071 [cs], November 2016. arXiv: 1611.03071.

[174] Sheila Jasanoff. Technologies of humility: citizen participation in governing science. Minerva, 41(3):223–244, 2003.

[175] Carlos Jensen and Colin Potts. Privacy Policies As Decision-making Tools: An Evaluation of Online Privacy Notices. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '04, pages 471–478, New York, NY, USA, 2004. ACM.

[176] Brian David Johnson. Science fiction prototyping: Designing the future with science fiction. Synthesis Lectures on Computer Science, 3(1):1–190, 2011.

[177] Mary Elizabeth "M.E." Jones, Melanie Kisthardt, and Marie A. Cooper. Interdisciplinary Teaching: Introductory Programming via Creative Writing. In Proceedings of the 42Nd ACM Technical Symposium on Computer Science Education, SIGCSE '11, pages 523–528, New York, NY, USA, 2011. ACM.

[178] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Rawlsian Fairness for Machine Learning. arXiv:1610.09559 [cs], October 2016. arXiv: 1610.09559.

[179] Justin Jouvenal. Police are using software to predict crime. Is it a holy grail or biased against minorities? Washington Post, November 2016.

[180] Immanuel Kant. Groundwork of the Metaphysics of Morals. Cambridge University Press, Cambridge, 1785.

[181] Fred Kaplan. WarGames and Cybersecuritys Debt to a Hollywood Hack. The New York Times, February 2016.

[182] Parija Kavilanz. Facebook 'likes' predict your personality, January 2015.

[183] Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. Standardizing privacy notices: an online study of the nutrition label approach. In Proceedings of the SIGCHI Conference on Human factors in Computing Systems, pages 1573–1582. ACM, 2010.

[184] Shalini Kesar. Including Teaching Ethics into Pedagogy: Preparing Information Systems Students to Meet Global Challenges of Real Business Settings. SIGCAS Comput. Soc., 45(3):432–437, January 2016.

[185] Youngsung Kim, ByungIn Yoo, Youngjun Kwak, Changkyu Choi, and Junmo Kim. Deep generative-contrastive networks for facial expression recognition. arXiv:1703.07140 [cs], March 2017. arXiv: 1703.07140.

[186] Youngsung Kim, ByungIn Yoo, Youngjun Kwak, Changkyu Choi, and Junmo Kim. Deep generative-contrastive networks for facial expression recognition. arXiv:1703.07140 [cs], March 2017. arXiv: 1703.07140.

[187] Jennifer King, Airi Lampinen, and Alex Smolen. Privacy: Is there an app for that? In Proceedings of the Seventh Symposium on Usable Privacy and Security, page 12. ACM, 2011.

[188] David Kirby. The Future is Now: Diegetic Prototypes and the Role of Popular Films in Generating Real-world Technological Development. Social Studies of Science, 40(1):41–70, February 2010.

[189] David A Kirby. Lab coats in Hollywood: Science, scientists, and cinema. MIT Press, 2011.

[190] Ben Kirman, Conor Linehan, Shaun Lawson, and Dan O'Hara. CHI and the future robot enslavement of humankind: a retrospective. In CHI'13 Extended Abstracts on Human Factors in Computing Systems, pages 2199–2208. ACM, 2013.

[191] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv:1609.05807 [cs, stat], September 2016. arXiv: 1609.05807.

[192] Enes Kocabey, Mustafa Camurcu, Ferda Ofli, Yusuf Aytar, Javier Marin, Antonio Torralba, and Ingmar Weber. Face-to-BMI: Using Computer Vision to Infer Body Mass Index on Social Media. arXiv:1703.03156 [cs], March 2017. arXiv: 1703.03156.

[193] Christine M. Korsgaard. Creating the Kingdom of Ends. Cambridge University Press, Cambridge, 1996.

[194] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15):5802–5805, April 2013.

[195] Adam DI Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. PNAS, 111(29):10779, 2014.

[196] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[197] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. Accountable Algorithms. SSRN Scholarly Paper ID 2765268, Social Science Research Network, Rochester, NY, March 2016.

[198] Mayank Kumar, Ashok Veeraraghavan, and Ashutosh Sabharwal. DistancePPG: Robust non-contact vital signs monitoring using a camera. Biomedical optics express, 6(5):1565–1588, 2015.

[199] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. arXiv:1607.02533 [cs, stat], July 2016. arXiv: 1607.02533.

[200] Erik Larson. The naked consumer: How our private lives become public commodities. Penguin Group USA, 1994.

[201] Jamiles Lartey. Predictive policing practices labeled as 'flawed' by civil rights coalition. The Guardian, August 2016.

[202] John Launchbury. A DARPA Perspective on Artificial Intelligence, 2017.

[203] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 34–42, 2015.

[204] Roger Eli Levien and ME Maron. A computer system for inference execution and data retrieval. Communications of the ACM, 10(11):715–721, 1967.

[205] Sam Levin and Olivia Solon Shaun Walker in Moscow. 'Our pain for their gain': the American activists manipulated by Russian trolls. The Guardian, October 2017.

[206] Tiancheng Li and Ninghui Li. On the Tradeoff Between Privacy and Utility in Data Publishing. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, pages 517–526, New York, NY, USA, 2009. ACM.

[207] Joseph Lindley. A pragmatics framework for design fiction. 2015.

[208] Joseph Lindley and Paul Coulton. Back to the Future: 10 Years of Design Fiction. In Proceedings of the 2015 British HCI Conference, British HCI '15, pages 210–211, New York, NY, USA, 2015. ACM.

[209] Joseph Lindley and Dhruv Sharma. Operationalising Design Fiction for Ethical Computing. SIGCAS Comput. Soc., 45(3):79–83, January 2016.

[210] Joseph Lindley, Dhruv Sharma, and Robert Potts. Anticipatory Ethnography: Design fiction as an input to design ethnography. In Ethnographic Praxis in Industry Conference Proceedings, volume 2014, pages 237–253. Wiley Online Library, 2014.

[211] Conor Linehan, Ben J. Kirman, Stuart Reeves, Mark A. Blythe, Joshua G. Tanenbaum, Audrey Desjardins, and Ron Wakkary. Alternate Endings: Using Fiction to Explore Design Futures. In CHI '14 Extended Abstracts on Human Factors in Computing Systems, CHI EA '14, pages 45–48, New York, NY, USA, 2014. ACM.

[212] Lorrie F. Cranor, Joseph Reagle, and Mark S. Ackerman. Beyond Concern: Understanding Net Users Attitudes about Online Privacy. In Ingo Vogelsang and Benjamin M. Compaine, editors, The Internet Upheaval: Raising Questions, Seeking Answers in Communications Policy, pages 47–70. MIT Press, Cambridge, Massachusetts, 2000.

[213] Ewa Luger, Stuart Moran, and Tom Rodden. Consent for All: Revealing the Hidden Complexity of Terms and Conditions. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, pages 2687–2696, New York, NY, USA, 2013. ACM.

[214] Kristian Lum and James Johndrow. A statistical framework for fair predictive algorithms. arXiv:1610.08077 [cs, stat], October 2016. arXiv: 1610.08077.

[215] Mary Madden and Lee Rainie. Americans' attitudes about privacy, security and surveillance. Pew Research Center, May 2015.

[216] Lydia Manikonda and Munmun De Choudhury. Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pages 170–181. ACM, 2017.

[217] Jennifer Mankoff, Jennifer A. Rode, and Haakon Faste. Looking Past Yesterday's Tomorrow: Using Futures Studies Methods to Extend the Research Horizon. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, pages 1629–1638, New York, NY, USA, 2013. ACM.

[218] Aaron Marcus. CHI at the Movies and on Tv. interactions, 13(3):54–ff, May 2006.

[219] Aaron Marcus. The History of the Future: Sci-fi Movies and HCI. interactions, 20(4):64–67, July 2013.

[220] Aaron Marcus. The Past 100 Years of the Future: HCI and User-experience Design in Science-fiction Movies and Television. In SIGGRAPH Asia 2015 Courses, SA '15, pages 15:1–15:26, New York, NY, USA, 2015. ACM.

[221] Ramia Maz and Johan Redstrm. Difficult forms: Critical practices of design and research. Research Design Journal, 1:28–39, 2009.

[222] Rene McCauley, Bill Manaris, David Heise, Cate Sheller, Jennifer Jolley, and Alan Zaring. Computing in the Arts: Curricular Innovations and Results. In Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, SIGCSE '17, pages 693–694, New York, NY, USA, 2017. ACM.

[223] Aleecia McDonald and Lorrie Faith Cranor. Beliefs and Behaviors: Internet Users' Understanding of Behavioral Advertising. SSRN Scholarly Paper ID 1989092, Social Science Research Network, Rochester, NY, August 2010.

[224] Gregory S. McNeal. Facebook Manipulated User News Feeds To Create Emotional Responses, June 2014.

[225] Jessica Mendoza. 'Predictive policing' isn't in science fiction, it's in Sacramento. Christian Science Monitor, August 2016.

[226] Michelle N. Meyer. Two Cheers for Corporate Experimentation: The A/B Illusion and the Virtues of Data-Driven Innovation. SSRN Scholarly Paper ID 2605132, Social Science Research Network, Rochester, NY, May 2015.

[227] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? Perspectives on Psychological Science, 6(1):3–5, January 2011.

[228] Jakub Mikians, Lszl Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. Detecting Price and Search Discrimination on the Internet. In Proceedings of the 11th ACM Workshop on Hot Topics in Networks, HotNets-XI, pages 79–84, New York, NY, USA, 2012. ACM.

[229] Claire Cain Miller. The Long-Term Jobs Killer Is Not China. Its Automation. The New York Times, December 2016.

[230] Claire Cain Miller. Evidence That Robots Are Winning the Race for American Jobs. The New York Times, March 2017.

[231] Lucia Moses. Marketers Should Take Note of When Women Feel Least Attractive, October 2013.

[232] Chantal Mouffe. Deliberative democracy or agonistic pluralism? Social research, pages 745–758, 1999.

[233] Omar Mubin, Mohammad Obaid, Wolmet Barendregt, Simeon Simoff, and Morten Fjeld. Science Fiction and the Reality of HCI: Inspirations, Achievements or a Mismatch. In Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction, OzCHI '15, pages 670–672, New York, NY, USA, 2015. ACM.

[234] Lewis Mumford. Technics and civilization. University of Chicago Press, 2010.

[235] Myriam Munezero, Calkin Suero Montero, Maxim Mozgovoy, and Erkki Sutinen. Exploiting Sentiment Analysis to Track Emotions in Students' Learning Diaries. In Proceedings of the 13th Koli Calling International Conference on Computing Education Research, Koli Calling '13, pages 145–152, New York, NY, USA, 2013. ACM.

[236] Kevin P Murphy. Machine learning: a probabilistic perspective. MIT press, 2012.

[237] Arvind Naranyanan, Joanna Huey, and Edward Felten. A precautionary appraoch to big data privacy. In Data protection on the move, pages 357–385. Springer Netherlands, 2016.

[238] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In 2008 IEEE Symposium on Security and Privacy (sp 2008), pages 111–125. IEEE, 2008.

[239] Arvind Narayanan and Shannon Vallor. Why Software Engineering Courses Should Include Ethics Coverage. Commun. ACM, 57(3):23–25, March 2014.

[240] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 427–436, 2015.

[241] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In Proceedings of the 23rd international conference on World wide web, pages 677–686. ACM, 2014.

[242] Rasmus Kleis Nielsen. Ground wars: Personalized communication in political campaigns. Princeton University Press, 2012.

[243] Max Nisen. Only 2% of Google's American Workforce Is Black. The Atlantic, May 2014.

[244] Jeff Nisker, Douglas K. Martin, Robyn Bluhm, and Abdallah S. Daar. Theatre as a public engagement tool for health-policy development. Health Policy, 78(2-3):258–271, October 2006.

[245] Helen Nissenbaum. Privacy as contextual integrity. Wash. L. Rev., 79:119, 2004.

[246] Helen Nissenbaum. Privacy in context: Technology, policy, and the integrity of social life. Stanford University Press, 2009.

[247] Martha C. Nussbaum. Sex and Social Justice. Oxford University Press, Oxford, 1999.

[248] William Odom, John Zimmerman, Scott Davidoff, Jodi Forlizzi, Anind K. Dey, and Min Kyung Lee. A Fieldwork of the Future with User Enactments. In Proceedings of the Designing Interactive Systems Conference, DIS '12, pages 338–347, New York, NY, USA, 2012. ACM.

[249] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: Privacy implications in social media. In European Conference on Computer Vision, pages 19–35. Springer, 2016.

[250] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA law review, 57:1701, 2010.

[251] GS Omenn, AC Kessler, NT Anderson, PY Chiu, J Doull, B Goldstein, J Lederberg, S McGuire, D Rall, and VV Weldon. Risk assessment and risk management in regulatory decision-making. Commission on Risk Assessment and Risk Management, 1996.

[252] Charles B. Owen, Laura Dillon, Alison Dobbins, Noah Keppers, Madeline Levinson, and Matthew Rhodes. Dancing Computer: Computer Literacy Though Dance. In Proceedings of the 14th International Conference on Advances in Mobile Computing and Multi Media, MoMM '16, pages 174–180, New York, NY, USA, 2016. ACM.

[253] Leysia Palen and Paul Dourish. Unpacking privacy for a networked world. In Proceedings of the SIGCHI conference on Human factors in computing systems, pages 129–136. ACM, 2003.

[254] Frank Pasquale. The black box society: The secret algorithms that control money and information. Harvard University Press, 2015.

[255] Frank Pasquale. The black box society: The secret algorithms that control money and information. Harvard University Press, 2015.

[256] Evan Peck. The Ethical Engine: Integrating Ethical Design into Intro to Computer Science, July 2017.

[257] Chanda Phelan, Cliff Lampe, and Paul Resnick. It's Creepy, But It Doesn'T Bother Me. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, pages 5240–5251, New York, NY, USA, 2016. ACM.

[258] Plato and Francis Macdonald Cornford. The republic of Plato, volume 30. Oxford University Press London, 1945.

[259] Christopher Plaue and Lindsey R. Cook. Data Journalism: Lessons Learned While Designing an Interdisciplinary Service Course. In Proceedings of the 46th ACM Technical Symposium on Computer Science Education, SIGCSE '15, pages 126–131, New York, NY, USA, 2015. ACM.

[260] Tomaso Poggio and Qianli Liao. Theory II: Landscape of the Empirical Risk in Deep Learning. arXiv:1703.09833 [cs], March 2017. arXiv: 1703.09833.

[261] James Poniewozik. Review: Black Mirror Finds Terror, and Soul, in the Machine. The New York Times, October 2016.

[262] Privacy Prosser. 48calif. L. REV, 383:389–92, 1960.

[263] John Pruitt and Jonathan Grudin. Personas: practice and theory. In Proceedings of the 2003 conference on Designing for user experiences, pages 1–15. ACM, 2003.

[264] Sarah Monisha Pulimood, Donna Shaw, and Emilie Lounsberry. Gumshoe: A Model for Undergraduate Computational Journalism Education. In Proceedings of the 42Nd ACM Technical Symposium on Computer Science Education, SIGCSE '11, pages 529–534, New York, NY, USA, 2011. ACM.

[265] Tarsem S. Purewal, Jr., Chris Bennett, and Frederick Maier. Embracing the Social Relevance: Computing, Ethics and the Community. In Proceedings of the 38th SIGCSE Technical Symposium on Computer Science Education, SIGCSE '07, pages 556–560, New York, NY, USA, 2007. ACM.

[266] Michael J. Quinn. Case-based Analysis: A Practical Tool for Teaching Computer Ethics. In Proceedings of the 37th SIGCSE Technical Symposium on Computer Science Education, SIGCSE '06, pages 520–524, New York, NY, USA, 2006. ACM.

[267] Rafael Ramirez, Malobi Mukherjee, Simona Vezzoli, and Arnoldo Matus Kramer. Scenarios as a scholarly methodology to produce interesting research. Futures, 71:70–87, August 2015.

[268] Rajib Rana. Gated Recurrent Unit (GRU) for Emotion Classification from Noisy Speech. arXiv:1612.07778 [cs], December 2016. arXiv: 1612.07778.

[269] Andrew G Reece and Christopher M Danforth. Instagram photos reveal predictive markers of depression. EPJ Data Science, 6(1), December 2017.

[270] The Nature of Respect. Stephen D. Hudson. Social Theory and Practice, 6(1):69–90, 1980.

[271] Maeve Reston. Voter data crucial to Romney's victory. Los Angeles Times, January 2012.

[272] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135–1144. ACM, 2016.

[273] Family Educational Rights and Privacy Act. USC 1232-34 CFR Part 99. 1974.

[274] Richard Rorty. Contingency, irony, and solidarity. Cambridge University Press, 1989.

[275] Richard Rorty. Philosophy and the Mirror of Nature. Princeton University Press, 2009.

[276] Rebecca J. Rosen. Why Today's Inventors Need to Read More Science Fiction. The Atlantic, September 2013.

[277] Mary Beth Rosson and John M. Carroll. Scenario-based Usability Engineering. In Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, DIS '02, pages 413–413, New York, NY, USA, 2002. ACM.

[278] Emma Ruby-Sachs. Amazon Censors Gay Books, May 2009.

[279] Christian Rudder. We Experiment On Human Beings!, July 2014.

[280] Kenneth John Ryan, JV Brady, RE Cooke, DI Height, AR Jonsen, P King, K Lebacqz, DW Louisell, D Seldin, E Stellar, and others. The Belmont Report. Washington, DC: US Department of Health, Education and Welfare, US Government Printing Office, 1979.

[281] Carl Sagan. The demon-haunted world: Science as a candle in the dark. Random House Digital, Inc., 1997.

[282] Koustuv Saha, Ingmar Weber, Michael L Birnbaum, and Munmun De Choudhury. Characterizing Awareness of Schizophrenia Among Facebook Users by Leveraging Facebook Advertisement Estimates. Journal of Medical Internet Research, 19(5):e156, 2017.

[283] Johnny Saldaa. The coding manual for qualitative researchers. Sage, 2015.

[284] Alton F. Sanders. A Discussion Format for Computer Ethics. In Proceedings of the 36th SIGCSE Technical Symposium on Computer Science Education, SIGCSE '05, pages 352–355, New York, NY, USA, 2005. ACM.

[285] Bruce Schneier. Essays: Lessons From the Dyn DDoS Attack - Schneier on Security.

[286] National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. Facilitating Interdisciplinary Research. The National Academies Press, Washington, DC, 2005. DOI: 10.17226/11153.

[287] Gargi Sharma. How Emotion Detection Technology Can Make Marketing Effective, April 2017.

[288] Katie Shilton. Participatory personal data: An emerging research challenge for the information sciences. Journal of the American Society for Information Science and Technology, 63(10):1905–1915, October 2012.

[289] Michael Shulman. Phones On, Curtain Up, July 2016.

[290] Karan Sikka, Gaurav Sharma, and Marian Bartlett. Lomo: Latent ordinal model for facial analysis in videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5580–5589, 2016.

[291] M. Skirpan and T. Yeh. Designing a Moral Compass for the Future of Computer Vision Using Speculative Analysis. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1368–1377, July 2017.

[292] Manya Sleeper, Justin Cranshaw, Patrick Gage Kelley, Blase Ur, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. I read my Twitter the next morning and was astonished: A conversational perspective on Twitter regrets. In Proceedings of the SIGCHI conference on human factors in computing systems, pages 3277–3286. ACM, 2013.

[293] Paul Slovic. Informing and educating the public about risk. Risk analysis, 6(4):403–415, 1986.

[294] Paul Slovic. Perception of Risk. Science, 236:280–285, April 1987.

[295] Paul Slovic. Perceived risk, trust, and democracy. Risk analysis, 13(6):675–682, 1993.

[296] Paul Slovic, Baruch Fischhoff, and Sarah Lichtenstein. Facts and fears: Understanding perceived risk. Societal risk assessment: How safe is safe enough, 4:181–214, 1980.

[297] Paul Slovic, Baruch Fischhoff, and Sarah Lichtenstein. Characterizing Perceived Risk. SSRN Scholarly Paper ID 2185557, Social Science Research Network, Rochester, NY, 1985.

[298] Debra L. Smarkusky, Stanley J. Stancavage, Ryan E. Eagan, Preston E. Propert, Raymond F. Plociniak, and Andrew M. Nichols. Physics in Motion: An Interdisciplinary Project. In Proceedings of the 2011 Conference on Information Technology Education, SIGITE '11, pages 33–38, New York, NY, USA, 2011. ACM.

[299] Debra L. Smarkusky, Sharon A. Toman, Peter Sutor, Jr., and Christopher Hunt. Performing Robots: Innovative Interdisciplinary Projects. In Proceedings of the 14th Annual ACM SIGITE Conference on Information Technology Education, SIGITE '13, pages 125–130, New York, NY, USA, 2013. ACM.

[300] Olivia Solon. Facial recognition database used by FBI is out of control, House committee hears. The Guardian, March 2017.

[301] Olivia Solon. 'This oversteps a boundary': teenagers perturbed by Facebook surveillance. The Guardian, May 2017.

[302] Carol Spradling, Leen-Kiat Soh, and Charles Ansorge. Ethics Training and Decision-making: Do Computer Science Programs Need Help? In Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education, SIGCSE '08, pages 153–157, New York, NY, USA, 2008. ACM.

[303] Aaron Springer, Victoria Hollis, and Steven Whittaker. Dice in the Black Box: User Experiences with an Inscrutable Algorithm. March 2017.

[304] Felix Stalder and Christine Mayer. The Second Index. Search Engines, Personalization and Surveillance. In Konrad Becker and Felix Stalder, editors, Deep Search: The Politics of Search Beyond Google, pages 98–115. Transaction Publishers, London, 2009.

[305] Jay Stanley. Does Surveillance Affect Us Even When We Cant Confirm Were Being Watched? Lessons From Behind the Iron Curtain.

[306] Bruce Sterling. Shaping things. MIT Press, Boston, MA, 2005.

[307] Bruce Sterling. Design Fiction. Interactions, 16(3):20–24, May 2009.

[308] S. Shyam Sundar, Hyunjin Kang, Mu Wu, Eun Go, and Bo Zhang. Unlocking the Privacy Paradox: Do Cognitive Heuristics Hold the Key? In CHI '13 Extended Abstracts on Human Factors in Computing Systems, CHI EA '13, pages 811–816, New York, NY, USA, 2013. ACM.

[309] Latanya Sweeney. Simple demographics often identify people uniquely. Health (San Francisco), 671:1–34, 2000.

[310] Latanya Sweeney. Information explosion. Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies, pages 43–74, 2001.

[311] Latanya Sweeney. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05):557–570, 2002.

[312] Latanya Sweeney. Discrimination in online ad delivery. Queue, 11.3:10, 2013.

[313] Joshua Tanenbaum. Design Fictional Interactions: Why HCI Should Care About Stories. interactions, 21(5):22–23, September 2014.

[314] Brian J. Taylor. Factorial Surveys: Using Vignettes to Study Professional Judgement. The British Journal of Social Work, 36(7):1187–1207, October 2006.

[315] Charles Taylor. The Politics of Recognition. In Multiculturalism: Examining the Politics of Recognition, pages 25–73. Princeton University Press, Princeton, 1992.

[316] Sam Thielman. Can we secure the internet of things in time to prevent another cyber-attack? The Guardian, October 2016.

[317] Sam Thielman. Your private medical data is for sale and it's driving a business worth billions. The Guardian, January 2017.

[318] Judith Jarvis Thomson. Killing, Letting Die, and The Trolley Problem. The Monist, 59(2):204–217, April 1976.

[319] Craig Timberg. Racial profiling, by a computer? Police facial-ID tech raises civil rights concerns., October 2016.

[320] Eran Toch, Yang Wang, and Lorrie Faith Cranor. Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. User Modeling and User-Adapted Interaction, 22(1-2):203–220, April 2012.

[321] Zeynep Tufekci. Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency. Colorado Technology Law Journal, 13.2:203–218, 2015.

[322] Zeynep Tufekci. Machine intelligence makes human morals more important, June 2016.

[323] Joseph Turow, Lauren Feldman, and Kimberly Meltzer. Open to Exploitation: America's Shoppers Online and Offline. A Report from the Annenberg Public Policy Center of the University of Pennsylvania, 2005.

[324] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. In Proceedings of the Eighth Symposium on Usable Privacy and Security, SOUPS '12, pages 4:1–4:15, New York, NY, USA, 2012. ACM.

[325] US Department of Health and Human Services. Protection of human subjects, 2009.

[326] Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. Machine Learning, 102(3):349–391, 2016.

[327] Elvira Perez Vallejos, Ansgar Koene, Chris James Carter, Ramona Statache, Tom Rodden, Derek McAuley, Monica Cano, Svenja Adolphs, Claire O'Malley, Kruakae Pothong, and Stephen Coleman. Juries: Acting out Digital Dilemmas to Promote Digital Reflections. SIGCAS Comput. Soc., 45(3):84–90, January 2016.

[328] Luise Vassie, Paul Slovic, Baruch Fischhoff, and Sarah Lichtenstein. Facts and fears: understanding perceived risk. Policy and Practice in Health and Safety, 3(sup1):65–102, 2005.

[329] Effy Vayena, Urs Gasser, Alexandra Wood, David R. O'Brien, and Micah Altman. Elements of a New Ethical Framework for Big Data Research. Washington and Lee Law Review Online, 72(3):420, 2016.

[330] James Vincent. Magic AI: these are the optical illusions that trick, fool, and flummox computers, April 2017.

[331] John Vines, Tess Denman-Cleaver, Paul Dunphy, Peter Wright, and Patrick Olivier. Experience Design Theatre: Exploring the Role of Live Theatre in Scaffolding Design Dialogues. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14, pages 683–692, New York, NY, USA, 2014. ACM.

[332] Jessica Vitak, Katie Shilton, and Zahra Ashktorab. Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community. pages 939–951. ACM Press, 2016.

[333] Joseph Voros. A primer on futures studies, foresight and the use of scenarios. Prospect: The Foresight Bulletin, 6(1), 2001.

[334] Nancy J. Wahl. YAATCEYet Another Approach to Teaching Computer Ethics. In The Proceedings of the Thirtieth SIGCSE Technical Symposium on Computer Science Education, SIGCSE '99, pages 22–26, New York, NY, USA, 1999. ACM.

[335] Daisuke Wakabayashi. As Google Fights Fake News, Voices on the Margins Raise Alarm. The New York Times, September 2017.

[336] Fulton Wang and Cynthia Rudin. Falling Rule Lists. In AISTATS, 2015.

[337] Yang Wang, Gregory Norice, and Lorrie Faith Cranor. Who is concerned about what? A study of American, Chinese and Indian users privacy concerns on social network sites. In International Conference on Trust and Trustworthy Computing, pages 146–153. Springer, 2011.

[338] Jeffrey Warshaw, Tara Matthews, Steve Whittaker, Chris Kau, Mateo Bengualid, and Barton A. Smith. Can an Algorithm Know the "Real You"?: Understanding People's Reactions to Hyper-personal Analytics Systems. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, pages 797–806, New York, NY, USA, 2015. ACM.

[339] Rick Wash. Folk models of home computer security. In Proceedings of the Sixth Symposium on Usable Privacy and Security, page 11. ACM, 2010.

[340] Kenji Watahiki and Motoshi Saeki. Scenario Evolution in Requirements Elicitation Processes: Scenario Pattern and Framework Approach. In Proceedings of the 4th International Workshop on Principles of Software Evolution, IWPSE '01, pages 166–169, New York, NY, USA, 2001. ACM.

[341] Wang Wei. Two Arrested for Hacking Washington CCTV Cameras Before Trump Inauguration.

[342] Joseph Weizenbaum. Computer power and human reason: from judgment to calculation. SH Freeman, 1991.

[343] Bernard Williams. The idea of equality. In Problems of the Self: Philosophical Papers 19561972, pages 230–249. Cambridge University Press, Cambridge, 1973.

[344] Langdon Winner. Do artifacts have politics? Daedalus, pages 121–136, 1980.

[345] Ursula Wolz and Lillian (Botos) Cassel. The Role of Interdisciplinary Computing in Higher Education, Research and Industry. In Proceedings of the 43rd ACM Technical Symposium on Computer Science Education, SIGCSE '12, pages 7–8, New York, NY, USA, 2012. ACM.

[346] Allen W. Wood. Kant's Ethical Thought. Cambridge University Press, Cambridge, 1999.

[347] Zoe J. Wood, Paul Muhl, and Katelyn Hicks. Computational Art: Introducing High School Students to Computing via Art. In Proceedings of the 47th ACM Technical Symposium on Computing Science Education, SIGCSE '16, pages 261–266, New York, NY, USA, 2016. ACM.

[348] Jenna Wortham. Finding Inspiration for Art in the Betrayal of Privacy. The New York Times, December 2016.

[349] Xiaolin Wu and Xi Zhang. Automated Inference on Criminality using Face Images. arXiv:1611.04135 [cs], November 2016. arXiv: 1611.04135.

[350] Xiaolin Wu, Xi Zhang, and Chang Liu. Automated Inference on Sociopsychological Impressions of Attractive Female Faces. arXiv:1612.04158 [cs], December 2016. arXiv: 1612.04158.

[351] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial Examples for Semantic Segmentation and Object Detection. arXiv:1703.08603 [cs], March 2017. arXiv: 1703.08603.

[352] David Jingjun Xu. The influence of personalization in affecting consumer attitudes toward mobile advertising in China. Journal of Computer Information Systems, 47(2):9–19, 2006.

[353] Yi Xu, Jared Heinly, Andrew M White, Fabian Monrose, and Jan-Michael Frahm. Seeing double: Reconstructing obscured typed input from repeated compromising reflections. In Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, pages 1063–1074. ACM, 2013.

[354] Hongyu Yang, Cynthia Rudin, and Margo Seltzer. Scalable Bayesian rule lists. arXiv preprint arXiv:1602.08610, 2016.

[355] Ke Yang and Julia Stoyanovich. Measuring Fairness in Ranked Outputs. arXiv:1610.08559 [cs], October 2016. arXiv: 1610.08559.

[356] Mu Yang, Yijun Yu, Arosha K. Bandara, and Bashar Nuseibeh. Adaptive sharing for online social networks: a trade-off between privacy risk and social benefit. In Trust, Security and Privacy in Computing and Communications (TrustCom), 2014 IEEE 13th International Conference on, pages 45–52. IEEE, 2014.

[357] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked Attention Networks for Image Question Answering. arXiv:1511.02274 [cs], November 2015. arXiv: 1511.02274.

[358] Yaxing Yao, Davide Lo Re, and Yang Wang. Folk Models of Online Behavioral Advertising.

[359] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. arXiv:1610.08452 [cs, stat], October 2016. arXiv: 1610.08452.

[360] Jinyan Zang, Krysta Dummit, James Graves, Paul Lisker, and Latanya Sweeney. Who knows what about me? A survey of behind the scenes personal data sharing to third parties by mobile apps. Proceeding of Technology Science, 2015.

[361] Kim Zetter. An Unprecedented Look at Stuxnet, the Worlds First Digital Weapon. WIRED, November 2014.

[362] Shanghang Zhang, Guanhang Wu, Joo P. Costeira, and Jos M. F. Moura. Understanding Traffic Density from Large-Scale Web Camera Data. arXiv:1703.05868 [cs], March 2017. arXiv: 1703.05868.

[363] Michael Zimmer. But the data is already public: on the ethics of research in Facebook. Ethics and Information Technology, 12(4):313–325, December 2010.

[364] Michael Zimmer. OkCupid Study Reveals the Perils of Big-Data Science, May 2016.

[365] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. Research through design as a method for interaction design research in HCI. In Proceedings of the SIGCHI conference on Human factors in computing systems, pages 493–502. ACM, 2007.

[366] Indre Zliobaite. On the relation between accuracy and fairness in binary classification. arXiv:1505.05723 [cs], May 2015. arXiv: 1505.05723.