ABSTRACT

Title of dissertation: UNDERSTANDING AND INTERVENING IN MACHINE LEARNING ETHICS: SUPPORTING ETHICAL SENSITIVITY IN TRAINING DATA CURATION

Karen Boyd, Doctor of Philosophy, 2020

Dissertation directed by: Professor Katie Shilton
University of Maryland School of Information

Despite a great deal of attention to developing mitigations for ethical concerns in Machine Learning (ML) training data and models, we don't yet know how these interventions will be adopted and used. Will they help ML engineers find and address ethical concerns in their work? This dissertation seeks to understand ML engineers' ethical sensitivity (ES)— their propensity to notice, analyze, and act on socially impactful aspects of their work—while curating training data. A systematic review of ES (Chapter 2) addresses conflicts of conceptualization in prior work by developing a new framework describing three activities (recognition, particularization, and judgment); argues that ES offers a useful way to describe, evaluate, and intervene in ethical technology development; and argues that the methods and perspectives of social computing can offer richer methods and data to studies of ES. A think aloud study (Chapter 3) tests this framework by using ES to compare engineers working with unfamiliar training data, finding that engineers with Datasheets noticed ethical issues earlier and more frequently than those without; finding that participants relied on Datasheets extensively while particularizing; and rendering rich descriptions of recognition and particularization in facial recognition data curation. Chapter 4 uses Value

Sensitive Design to "design up," mitigating harms by helping machine learning engineers particularize their ethical concerns and find appropriate technical tools. It introduces ES to studies of social computing, contributes a novel method for studying ES, offers rich data about how it functions in ML development, describes insights for designing context documents and other interventions designed to encourage ES, develops an extensible digital guide that supports particularization and judgment, and points to new directions for research in ethical sensitivity in technology development.

UNDERSTANDING AND INTERVENING IN MACHINE LEARNING ETHICS: SUPPORTING ETHICAL SENSITIVITY IN TRAINING DATA CURATION

by

Karen Boyd

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Katie Shilton, Chair/Advisor
Professor Hal Daumé III
Professor Wayne Lutters
Professor Jessica Vitak
Professor Susan Winter

Dedication

To failure, my least favorite and most effective teacher.

Acknowledgments

This all still feels quite impossible, but surely it would have been without immense professional and personal support.

First, thank you to Dr. Katie Shilton, who took a chance on me in the first place. I couldn't have asked for a better advisor or example of how to be a scholar. Thank you for your patient support and guidance.

Thank you to the faculty at the University of Maryland, especially everyone who served on committees– Dr. Hal Daumé III, Dr. Wayne Lutters, Dr. Yla Tausczik, Dr. Jessica Vitak, and Dr. Susan Winters– for your clear and constructive feedback, flexibility, and support of my work. In addition, I was especially encouraged by Dr. Kari Kraus, Dr. Katy Newton Lawley, and Dr. Vedat Diker who offered me encouragement and mentorship outside the research arena; your help will help me make a better, broader impact.

Thank you to my participants for your time and thoughts– it was invaluable and I treasure your trust. The San Diego Machine Learning Meet-up especially helped connect me with the community, continues to improve my fluency in the technology and language of ML, and occasionally gets in very amusing fights about self-driving cars or unsupervised methods.

Thank you to Brandon Smith for serving as an key informant, expert programmer, and unfailing personal support. When we met, I was a "failure" and now I am a doctor. The first moment I would have quit without your encouragement was before I applied, and my attempts to sabotage myself were pretty non-stop from there. You've put up with a great deal of my stress and stubborn self-doubt over the last 5 years; I can't begin to thank you appropriately for your persistent help and belief in me.

Thank you to my family, especially to my parents. You raised me to ask questions, to get to the bottom of things, and to adapt to change. You also believed I could do better, and

you were right. I hope I can see students' potential through their difficulties the way you did for me. Andy, thanks for stirring up conversation and for more tacos, burgers, beer, and wine than I should have ever consumed[1].

Thank you to Rupal Mehta for believing in me always, encouraging me constantly, and showing me how it's done. Thank you to the Consortium for the Science of Sociotechnical Systems Researchers for introducing me to my research family; special thanks to Stevie Chancellor, Michael Ann DeVito, Michaelanne Dye, Sarah Gilbert, Elliot Hauser, Koko Koltai, and Jacob Thebault-Spieker who remind me I belong here (and make it much more fun). Thanks to Cassidy Callahan, Verna Seth Clark, Rebecca Daniel Sigmon, Brittany Smith, and Luke Valles for keeping me entertained and engaged in the world outside.

I read on the Internet that it's rude to thank your cat before funding bodies and actual people who can read, so Mandelbrot, this part is for you. You're a great cuddler, a superb practice talk audience, and have been, for the most part, a peaceful presence through an especially trying time. Thank you to the Humane Rescue Alliance for caring for pets in the Washington DC area, especially one cat formerly known as "China Doll."

[1] #worthit

Table of Contents

Chapter 1: Introduction

As machine learning (ML) techniques have become sophisticated and pervasive, ethical concerns have followed. Recent headlines have declared "Amazon scraps secret AI recruiting tool that showed bias against women," "Google Photos labeled black people 'gorillas"' and "U.S. charges Facebook with racial discrimination in targeted housing ads" [50, 83, 141]. Alongside journalists, researchers have verified algorithmic discrimination in outcomes and accuracy on the basis of age [52], gender [25, 50], race [14, 124, 171], and the intersection of gender and race [33, 135], across product types, including (for the above examples) text processing, search engines, facial recognition, ad delivery, and criminal risk estimates.

Fairness isn't the only ethical implication of ML: concerns about privacy and accountability have also been raised. Publicly released datasets said to be anonymized were re-identified [122, 132, 169] and researchers, the popular press, and the courts are discussing algorithmic accountability and due process in big data [7, 14, 49, 156].

In response to ethical concerns, researchers have designed, tested, and published technical and practice-based interventions throughout the ML development process. Conferences like ACM's Fairness Accountability and Transparency (FAccT) and Artificial Intelligence, Ethics, and Society (AIES) focus on fairness, accountability, and other values across ML techniques. Centers like AI Now and The Institute for Ethical AI and Machine Learning are built around responsible artificial intelligence (AI), and existing centers, like Berkman Klein, have added it as a topic of focus.

Despite increasing attention to intervening in ML development, we don't yet know whether interventions designed to promote ethical AI will work to help ML engineers recognize, understand, and make effective decisions about ethical issues in their work. Holstein et al. made a strong argument that tools to mitigate ethical concerns in ML development must be designed with the experiences, practices, and perspectives of ML engineers in mind [93]. My dissertation seeks to understand ML engineers' ethical sensitivity– their propensity to notice, analyze, and act on socially impactful aspects of their work– and how it influences their use of an intervention designed to mitigate ethical concerns during training data curation. It is composed of three papers. The first reviews the literature on ethical sensitivity across disciplines and links it to existing work in social computing. The second describes a study using ethical sensitivity to examine the effectiveness of a proposed intervention into the practices of ML engineers during training data curation, specifically whether it helps them recognize ethical issues. The third uses value sensitive design to develop a novel intervention that may help ML engineers build a thorough understanding of an ethical issue and link their understanding to a judgment.

1.1 Bias in Machine Learning

Machine learning is a broad and contested category of techniques. Generally, ML techniques fall into three categories: supervised, unsupervised, and reinforcement learning. Supervised learning methods, like neural networks (which includes the much-discussed deep learning), are given training data that can include millions of examples (rows of data), several characteristics of each example, and a label, which operates like a correct answer. The algorithm then builds itself to predict the "correct answers" for each example with as little error as possible, and the resulting model is used to predict answers for new examples. In contrast, unsupervised methods (like clustering algorithms or anomaly detection)

find patterns in data that are not centered around a single particular label, but are instead groups of examples that have patterns of characteristics in common. Unsupervised methods can be used to group songs that have features in common, customers who behave similarly, and other clustering problems, for example. In reinforcement learning, a reward function is used to teach the algorithm how to achieve desired goals, like driving an autonomous vehicle. There are some "semi-supervised" methods and some groups of methods, like Natural Language Processing (NLP), that incorporate techniques from more than one category.

One thing ML algorithms have in common is a requirement for input data: a set of data, usually large, that mirrors what its creators want to predict or estimate and from which the algorithm will deduce patterns [188]. Although there's work being done to improve algorithmic transparency [55, 152], many ML techniques in use do not result in algorithms that are interpretable by humans– even by the people who built them [34].

Society often imagines that giving weighty societal problems over to computers will circumvent human weaknesses, like inconsistency, prejudice, or limited processing speed [17]. In the case of machine learning, algorithms may be able to make consistent predictions quickly, but they still learn from data collected from a world where human cognitive, social, and institutional biases are at work. For example, ML is used because it can alleviate backlogs in social systems like the parole process by making decisions quickly and it can eliminate inconsistency based on the circumstances of a particular judgment decision, for example, the mood of a judge on a particular day [111]. While ML algorithms can smooth out this kind of judge-to-judge and day-to-day inconsistency, appearing to be more fair, ML techniques do not eliminate systemic bias, including social prejudice and inequality. In fact, these algorithms learn and perpetuate systemic bias at scale [17, 135]. Algorithms can't distinguish between fair and unfair patterns in the data: they can only detect patterns that improve predictions. So, if a certain jurisdiction has a history of race-based discrimination and wants to create a risk assessment algorithm for parole, the algorithm will learn

that the feature "race" (or correlates of race, like zip code) is an useful predictor of the parole board's decision and will use it to evaluate re-offense rate, resulting in parole decisions that are biased by race [111, 136].

1.2 Interventions

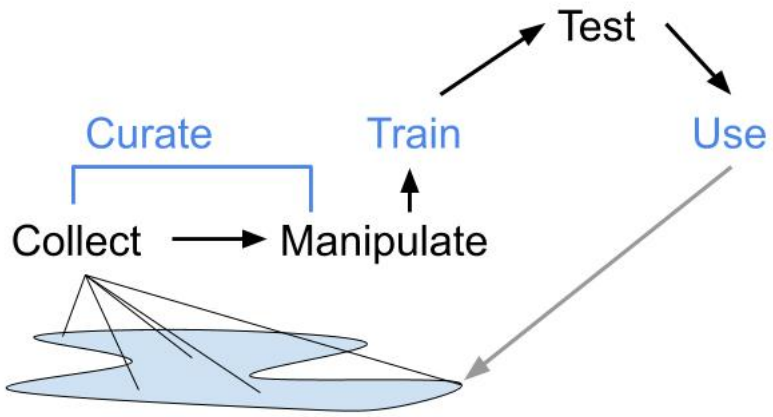
Researchers in academia and industry have developed a wide range of interventions that ML engineers can use to address fairness, privacy, and accountability. These tools can be classified in a couple ways, including where in the life cycle of training data they intervene and how.

1.2.1 Machine Learning Lifecycle

Machine learning is a very diverse set of methods, but in general, teams must collect or aggregate large sets of training data. Engineers then will do some data cleaning and manipulation with the goal of making the data consistent and as close to the case in which the algorithm is intended to be used as possible. They may oversample important rare classes to ensure that they can be detected, undersample majority cases, manipulate existing examples to create more data (for example, by mirroring images), or otherwise ensure that the training data is sufficient. I will refer to collection, cleaning, and manipulation together as training data curation.

Engineers define the task, asking "what should the goal of this algorithm be?" What is it predicting? Classifying? What cost function should it be minimizing? What kinds of errors are heavily and less heavily penalized? One of several techniques will be selected to automatically find patterns in the data that can be used to predict or classify future data. The resulting algorithm may be used on its own, assembled with others, or used as part of other software. Algorithms and the products that drive them are tested, retrained or adjusted if

Figure 1.1: Training Data Resourcing Cycle



necessary, and then released for use. Often, data from users' behavior is added to the pool of potential data and is reused as training data for this or another system (see Figure 1.1).

Interventions that aim to change the ethical impact of machine learning models are often classified by when they intervene: before the model has been trained (by manipulating pre-processed training data), during training (by modifying the algorithm), or after training (by adjusting the algorithm's outputs) [70]. These moments for intervention are labeled in blue in Figure 1.1.

Pre-processed training data has been identified as a driver of discrimination. It reflects biases that exist in the world from which the data was drawn, for example in associations between words in text [37] or demographics and hiring [140]. Training data has also become a target for accountability interventions, like Datasheets, the dataset nutrition label, and data statements for NLP [21, 74, 91] and is described by industry practitioners as a key place to intervene to support fairness in ML [93]. Data collection is an area of particular concern for privacy advocates, but the issue is complicated because collecting sensitive attributes may be necessary to build and certify fair algorithms [183]. Therefore, researchers advocate interventions into training data to preserve privacy and allow for fair model training, model certification, and decision verification by, for example, encrypting sensitive

attributes in training data [108]. Because training data is cited as a driver of discrimination, a focus of concern for multiple ethical issues, and highlighted by industry practitioners as a target for intervention, this project focuses on pre-processed training data.

1.2.2 Practice and Product Interventions

This dissertation also classifies interventions in machine learning based on whether they intervene in the products of ML development– through technical interventions– or in the practices of ML development. Interventions can take place in either practices or products at any stage of the training data resourcing cycle. For example, a technique for manipulating training data to promote fairness is a pre-processing intervention into the product of training data, while a checklist to ensure that training data is representative of minority classes is a intervenes before processing into the practices around training data curation.

Computer scientists have produced many technical interventions. These are techniques to alter the products of ML work– including training data, learning algorithms, and ML models– in order to address ethical concerns. Most of these are limited in their scope: they can be applied to a single ML technique, a single ethical issue, or to protect a single, known, and well-defined demographic group. For example, they may aim to achieve similar error rates for two populations in college admissions or credit approval [20] which is appropriate if you already know that disparity exists and there are two, clearly delineable groups, but not if there are three demographic groups with ambiguous or overlapping membership, if there are privacy concerns as well as fairness ones, or if the decision-support system comprises more than one algorithm. These bounds are not fatal flaws; in fact, they allow technical interventions to effectively target known issues in training data.

Interventions into the practice of developing ML may help guide engineers to identify the need for technical interventions and guide them to select appropriate ones. For

example, using an ethics checklist (e.g., "Deon"[1]) or toolkit designed to assess and manage risks of harm (e.g., "Ethics & Algorithms Toolkit" [2]) may raise relevant ethical issues early in development and allow teams to select the most effective technical intervention, instead of having to find a way to alter a nearly-complete product or read about harms the algorithm has caused in headlines. This dissertation will focus on practice interventions intended to promote ethical sensitivity in ML engineers as they curate training data: Chapter 3 will evaluate one such intervention and Chapter 4 will develop a practice intervention that guides users to select appropriate technical ones.

1.3 Theoretical Framework

This dissertation relies on the concept of ethical sensitivity to understand the way that ML engineers notice, perceive, and judge ethical concerns in their work. Ethical sensitivity has been used to describe the work of professionals who make ethically consequential decisions but do not have concrete heuristics to determine whether a decision is ethically consequential, including nurses, accountants, teachers, doctors, and marketers [186].

Ethical sensitivity describes a professional going about the technical aspects of their work until they notice a cue that causes them to *recognize* a potential ethical issue. They then *particularize* the ethical issue: they use the situational and external information they have to estimate the scope, scale, and details of the issue and determine how it relates to their norms, expectations, and job task. They will then make a *judgment* about whether and how to act.

This study will therefore seek to identify cues, describe the circumstances of ethical recognition, and describe particularization and judgment. Each of these concepts are described in detail in the Ethical Sensitivity in Machine Learning section of Chapter 2 and

[1] https://deon.drivendata.org/
[2] https://ethicstoolkit.ai/

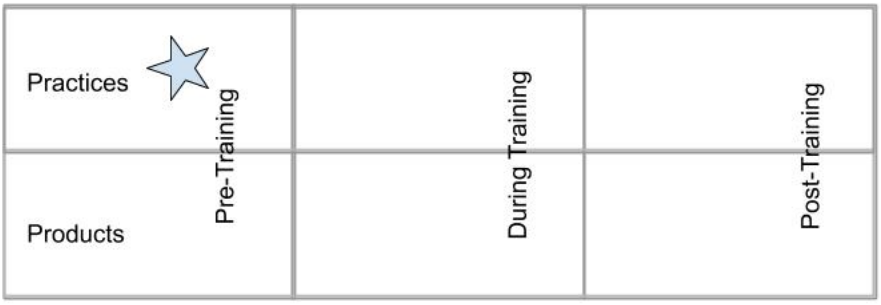


Figure 1.2: Ways to Intervene in Machine Learning

operationalized in the Data Collection section of Chapter 3. The goal of the study design will be to arrange an opportunity to observe ethical recognition, particularization, and judgment as ML engineers explore new data in a situation that is similar to their normal working environment.

The results of this study will be rich description of how ML engineers explore unfamiliar datasets; how they recognize, particularize, and judge; and how context documents and ethical guides may affect ethical sensitivity. These findings can be used to improve existing documents and inform the development of new tools.

1.4 Scope

Section 1.2.1 described 6 ways of intervening in ML: before, during, or after training, targeting the products or practices of ML work. This project will focus on interventions into ML development practices before training occurs, indicated by a star in Figure 1.2.

The interventions I will test assume that engineers will need to identify an ethical concern, understand how it relates to the specifics of their model and task, select means of mitigating or managing the issue, and then implement and test it: the activities described by ethical sensitivity. This dissertation will look at the role of context documents in per-

ceiving and particularizing an ethical issue and of a mitigations guide in moving from particularization to judgment.

1.4.1 Context Documents

It's often the case that the same team collects training data and prepares it to train a model, but it is often not. It may be different teams' responsibility in a large organization, they may reuse data collected for other purposes (for example sales, quality control, or user data), and they may use any of many large, public datasets available. OpenML lists more than 2,600 open datasets [178].

Recently, researchers have called for standard documentation accompanying datasets or ML-driven systems that describe characteristics of training data and help others interpret and use the data or model– a type of practice intervention I'll call "context documents." They function as a way to communicate information about data provenance from those who collect it to those who use it. Section 3 will describe this disconnect in more detail. These context documents take many forms, ranging in complexity from a few hundred words [156] to detailed reports [21, 91]. Proposals like Bender and Friedman's [21] for Natural Language Processing and Yang et al.'s for ranking algorithms [193] illustrate the specificity that context documents tailored for a single ML technique can offer. Some are part of larger programs or regulatory regimes and have a format tailored to their purpose in it [146, 158, 166]. Gebru et al.'s Datasheets [74], Mitchell et al.'s Model Cards [127], and Yang et al.'s nutritional label [193] directly ask for information about ethical concerns, while others argue that simply reporting the characteristics of datasets will prompt and advertise ethical work [21, 91]. The sudden proliferation of context document proposals may be a response to an uptick of research and journalism verifying algorithmic bias: all but one of these context document proposals cited either Julia Angwin's "Machine Bias"

[14], Bolukbasi et al.'s "Man is to Computer Programmer as Woman is to Homemaker?" [25], or both.

Chapter 3 describes a study focused on Datasheets [74]: a technique- and domain-agnostic, lay-language context document for training data. Datasheets are versatile: they can be taught early in ML education to students who will go on to work in diverse domains using a variety of techniques and can be read by non-experts, like managers, users, citizens, and auditors.

1.4.2 Mitigation Guide

In response to calls for attention and action for ethics in ML, particularly around fairness, some repositories of mitigation techniques have been developed. Instead of searching through scholarly articles, which they may not have access to, engineers can visit, for example, Intel's list of AI Ethics toolkits list [1] or IBM's AI Fairness 360 Open Source Toolkit, which includes tutorials and code for several bias mitigation algorithms [2].

These projects appear to be useful, especially in that they offer checklists, tutorials, and code, which are out of the scope of the mitigation guide that this dissertation offers. However, in order to find an appropriate mitigation they require engineers to already know a lot of characteristics of their problem and its solution. In other words, they need to have already particularized the problem quite a bit. Take for example the IBM resource. The page lists the titles of algorithms, a sentence about when in the process it is used, and a sentence about what it changes. For example, "Adversarial Debiasing: Use to mitigate bias in classifiers. Uses Adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions" or "Reject Option Classification: Use to mitigate bias in predictions. Changes predictions from a classifier to make them fairer." These descriptions are useful if the engineer already knows when they want to intervene and how.

The goal of the mitigation that this dissertation produced is to enable search by problem characteristics to aid in particularization as well as judgment. For example, an engineer has identified that their dataset includes a minority or protected class that is poorly represented. Development of this tool started with the assumption that, rather than being organized by solution characteristics (where it intervenes and how), engineers may need a tool that is organized by problem characteristics: does the algorithm produce different error rates for these groups? Different predictions? Are there two groups or could there be several?

Chapter 4 describes a Value Sensitive Design study that developed a mitigation guide with the goal of supporting particularization to judgment. Scaffolding ethical sensitivity may make it easier for engineers to practice it in their work.

1.5 Structure of Project

This dissertation project has a systematic literature review (Chapter 2) and two empirical phases, each designed to move engineers through different parts of ethical sensitivity. Although it's possible and likely that many individual engineers will begin judgment earlier or decide that the ethical issue is not relevant to their job at all, in general, the document in Chapter 3 is designed to prompt perception and improve particularization and the document in Chapter 4 will be designed to move participants from particularization to judgment.

1.5.1 Testing Datasheets

The first empirical phase of this project describes whether and how Datasheets are used by model builders in the early stages of data exploration and takes steps toward evaluating Datasheets' effectiveness in encouraging ethical sensitivity, particularly whether they cue recognition of ethical issues and how ML engineers use them to particularize. It will contribute to literature exploring the role of context documents in technical work, further

the effort to create effective context documents for training data, and inform the development of training, best practices, and other kinds of interventions into ethical discovery and training data exploration.

Chapter 3 will explore how introducing Datasheets may change ML engineers' approach to ethics in a machine learning problem:

> **RQ 1**: Are engineers who read Datasheets cued to perceive ethical problems differently or at different rates than those who do not read them?
>
> **RQ 2**: What information on and off the Datasheet do ML engineers use to particularize a perceived ethical problem?

1.5.2 Value Sensitive Design

The second empirical phase of this project designs a living mitigation guide to help ML engineers particularize and make judgments about known ethical issues. An open source taxonomy of diagnostic tools and ethical interventions may help spark ideas, empower ML engineers to see that they can mitigate ethical issues in ML, raise awareness of key aspects of ethical problems in ML, improve particularization skills, and reduce the time and effort required to select mitigations. Such a document may also be used as a tool for training and research. This phase uses Value Sensitive Design (VSD) [71] to address the following design goals:

1. Enhance users' ability to particularize and judge ethical problems in training data.
2. Improve awareness of existing and new technical interventions among practitioners and researchers
3. Empower trainers, educators, and leaders in ML with structured and restructurable information about technical interventions for ethical concerns in training data

4. Achieve above design goals while minimizing interruption to ML engineers' work practices

1.6 Empirical Methods

This dissertation employs a mixed methods approach in order to get a detailed picture of ethical sensitivity in ethical ML development. First, a comparative think-aloud study allows us to observe recognition and particularization among machine learning engineers who are considering unfamiliar and ethically-complicated training data. Second, a value sensitive design study uses this particularization data along with additional, guided particularization data and targeted interview questions to develop a guide to ethical mitigation guide for ML engineers.

1.6.1 Chapter 3: Think Aloud Experiment

Chapter 3 describes ethical sensitivity while reviewing and evaluating unfamiliar and ethically-fraught facial recognition training data, both with and without a Datasheet. The study uses audio and screen recordings of participants exploring a dataset and speaking aloud as they do so; some participants were given Datasheet describing the provenance and characteristics of the dataset, and some were not. The output of this study is detailed data about the impetus and nature of ethical recognition and particularization; whether, what, and how much information in the Datasheet is referred to by the engineers as they particularize; and qualitative information about how participants work with unfamiliar data with and without a Datasheet. Information from this study can inform the refinement and development of interventions supporting ethical recognition and particularization in ML development.

1.6.2 Chapter 4: Value Sensitive Design

After observing and listening to each participant as they worked with an unfamiliar dataset to solve an ML problem, I interviewed them about their next steps, perceptions of the ethical problem, and perceptions of ethics in their own work. Then, they were briefed about ethical issues in the facial recognition data and asked to assist with some portion of the value sensitive design study.

I iterated among conceptual, technical, and empirical investigations as prescribed by VSD. Conceptual investigations included a literature review and stakeholder analysis. Empirical investigations used particularization data from the think aloud, direct questions about particularization habits and ethical experience, and think-aloud data as they sought ethical mitigations for a performance bias problem in the facial recognition data. I worked with a professional programmer for the technical investigation to produce a prototype tool[3] and tested it with ML engineers and technology ethics researchers.

1.7 Importance and Contribution

This work will answer questions about the potential impact of context documents and ethical guides in the development of ML-driven systems, explore and operationalize ethical sensitivity in a new and consequential profession, offer a new method for studying ethical sensitivity, richly describe ML development practices at a key stage, develop a tool to help ML engineers (and managers and educators) particularize and judge ethical problems in ML, and offer guidance for intervening in training data curation.

First, it helps understand how best to design and deploy context documents. Specifically, this work offers data about whether a Datasheet aids ML engineers in recognizing, particularizing, and judging an ethical issue; reveals how to make existing and new context

[3] https://ml-ethics-tool.web.app/

documents more effective; and suggests training, business processes, and other scaffolds to improve the usability and effectiveness of context documents. These results offer guidance for all key players in the ML-driven systems ecosystem: data curators, model builders, model users, and, if model builders and data curators choose to publish these documents, auditors, regulators, and citizens.

Second, it designs and evaluates a tool that may help ML engineers explore and select interventions to mitigate known ethical issues highlighting key aspects of problem-mitigation fit; raise awareness of ethical issues, values, opportunities and ways to intervene; and make interventions for (and make arguments for intervening in) a given ethical issue in a given circumstance easier to find. This document may also be useful for training ML engineers and exposing areas in need of research and development of new interventions.

Third, it introduces and starts to illuminate ethical sensitivity in technology development, and Machine Learning development in particular. It introduces, conceptualizes, and operationalizes ES for use in technology development. It offers a method for testing interventions aimed at encouraging recognition, supporting particularization, and guiding judgment in technology development. It highlights information types and sources ML engineers rely on when particularizing and describes the ethical sensitivity of ML engineers faced with an unfamiliar, ethically-fraught dataset.

The long-term goal of improving document interventions is to encourage the ethical development of ML and AI systems. Some of the most immediate harms from pervasive ML are bias issues that have consistently targeted underrepresented groups. Encouraging ethical development of ML will directly benefit those who are currently under-served or harmed by these emerging technologies.

This study is designed to improve an existing ethical intervention into ML development practices and propose a new one. It does not endorse or offer a single intervention that can be deployed in all contexts. This dissertation can be a jumping-off point for machine

learning teams, managers, and researchers who want to better understand how to effectively integrate these documents into their workflow.

1.8 Organization of the Dissertation

The following sections will include three papers: a systematic literature review entitled "Ethical Sensitivity: Advancing Methods for Studying Ethics in Technology Development" co-authored with my advisor, Katie Shilton (Chapter 2); a paper presenting the context document study entitled "Datasheets for Datasets help ML engineers notice and understand ethical issues in unfamiliar training data" (Chapter 3) and a paper presenting the VSD study entitled "Designing Up with Value-Sensitive Design: Building a Guide for ethical machine learning development" (Chapter 4) The final section will summarize the findings of these papers, their contributions, weaknesses, and next steps.

Chapter 2: Ethical Sensitivity: Advancing Methods for Studying Ethics in Technology Development

This chapter is an article under review. It was co-authored with Dr. Katie Shilton. I compiled and analyzed the corpus, synthesized the framework, conceptualized and organized the paper. Dr. Shilton contributed background knowledge of cooperative work and social computing to help apply the framework.

Abstract Studying technologists' engagement with the ethical aspects of their work is important, but engagement with ethical issues is an unobservable construct without agreement on what observable factors comprise it. Ethical sensitivity (ES), a construct studied in medicine, accounting, and other professions, offers a framework of observable factors by operationalizing ethical engagement in workplaces into component parts. This paper uses a corpus of 108 ES studies from 1985-2020 to adapt the framework for research in social computing ethics. We use this literature to build an umbrella framework that conceptualizes ES as including the moment of noticing an ethical problem (recognition), the process of building understanding of the situation (particularization), and the decision about what to do (judgment). This framework makes theoretical and methodological contributions to the study of how designers consider ethics. We find that ethical sensitivity can provide useful language for describing studies in technology ethics; suggests opportunities for, and evaluations of, ethical interventions for design workplaces; and can help researchers connect individual backgrounds, educational experience, work practices, and occupational and or-

ganizational factors to design decisions. Simultaneously, existing social computing ethics research methods can expand the limited range of research methods currently employed in the current ES literature, adding rich, contextualized data about ethics in work practice.

2.1 Introduction

How technologists consider ethics during design and development has become a matter of public concern and substantial research. Increasingly, researchers and the public have demanded that ethical reflection become a central part of computing. Current research in HCI and CSCW engages with this challenge by studying how technologists 'do' ethics during design, framing ethics as a team practice in the tradition of studying design work practices [30, 73, 99, 101, 192]. Some CSCW and HCI work has studied situated aspects of design that impact values and ethical implications. For example, Chivukula et al. [43] found five organizational and practice-based dimensions that influence designers' ethical awareness and understanding, ranging from the positionality of the design task within the enterprise to designer and stakeholder education. Shilton [163, 164] has described work practices that surface ethics discussions on design teams as values levers. And many HCI and CSCW researchers have created artifacts and interventions to encourage technology designers to notice and engage with ethical issues in their work [65, 162]. For example, Wong et al. [192] and Baumer et al. [18] have used design fiction to elicit "contextual, socially-oriented understandings" of values and ethical implications of potential features among designers. Value-sensitive design methods [71] and card sets [12, 72, 118] evoke, engage, or elicit values during design, helping designers to make their values concrete. Gispen offers tools and exercises designed to develop moral sensitivity, moral creativity, and moral advocacy among designers [76].

Ethics as a professional practice, with or without outside intervention, has also been

studied extensively outside of computing. In particular, the framework of *ethical sensitivity* (ES) focuses on how professionals recognize, interpret, and make ethical decisions in their work [184]. The ES framework lends itself to studies of practice by focusing on the processes that scaffold ethical decision-making: research that asks not if professionals are ethical, but how and when they engage in ethical practices and decisions, and how work might be adapted to encourage ethical practices. And because there is a well-developed literature on ethical sensitivity in professions outside of computing, this framework has the advantage of clear conceptual definitions: both shared indicators of the otherwise unobservable construct of ethical sensitivity, and methods to study those indicators. We draw on previous conceptual analyses of ES as well as a systematic review of empirical ES studies to condense the wide range of proposed attributes of ethical sensitivity into three umbrella components: recognition, particularization, and judgment (Figure 2.1).

First, workers experience recognition of a potential ethical issue in their daily tasks (perhaps prompted by an external cue or internal affectivity). Workers then, individually or in teams, particularize the details of the situation through reflection (e.g., on their personal values) or information seeking (e.g., finding external guidance such as codes of ethics or opinions of peers and managers). Finally, workers use the understanding they build to make judgments about what to do. Recognition, particularization, and judgment offer defined activities to observe and measure for HCI and CSCW researchers interested in evaluating ethics interventions, describing a team's ethical practices, or understanding the impact of design complexity, organizational factors, and other variables of interest on design decision-making. The ES framework can help CSCW and HCI researchers systemize both where to look for ethical practices, and how to test interventions into the ethical practices of technology designers.

Our review of the methods and indicators widely used to study ES also reveals two important limitations of the existing ES literature for studying ethics in technology design.

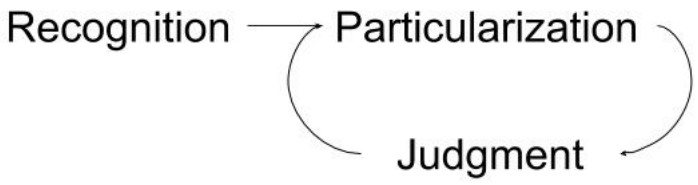


Figure 2.1: Ethical Sensitivity stages

First, ES has been studied almost exclusively as a static property of individuals, rather than a skill that can be developed or a feature of teams. Second, the current methods that dominate the ES literature neglect aspects of work practice particularly important within HCI and CSCW: the situated and embodied nature of technology design work [57, 129, 170]. In addition to adapting and refining a useful framework for studying ethics during the team practices of technology design, this paper highlights opportunities for HCI and CSCW research to offer rich, situated data from contextual methods to the study of ES. Our paper proceeds as follows: Section 2.2 describes our systematic review of the ES literature. Section 2.3 describes how we condensed the many concepts that make up ES into a framework for studying ethics in technology design. Section 2.4 then explores the methods and indicators most frequently used to study the three umbrella components of ES. Section 2.5 returns to the HCI and CSCW literature, using existing studies of ethics work during technology design to discuss the limitations of existing ES methods for researchers from HCI traditions and suggest how existing literature in HCI and CSCW enriches the ES literature. Section 2.6 discusses recommendations for social computing researchers interested in employing ES as a framework to guide their own research.

2.2 Constructing the ES Corpus

We began our analysis of the ES literature with review articles [24, 88, 184–187] to better understand the history and context of the theoretical framework and to develop a

search vocabulary for our systematic literature review. The review articles revealed that: 1) ethical sensitivity has been studied in numerous disciplines; 2) what is now called "ethical sensitivity" was originally conceptualized and is sometimes still studied as "moral sensitivity;" and 3) existing review articles did not include an explicit discussion of indicators or methods used to operationalize or measure ethical sensitivity.

Based on our analysis of the review articles, we searched Google Scholar for papers in English containing either "ethical sensitivity" OR "moral sensitivity," returning 17,400 papers that used either phrase. Next, we limited the corpus to empirical studies whose focus was the phenomenon in question: excluding papers that referred to "ethical sensitivity" in their text but didn't attempt to measure it. We also excluded studies about ethical judgment alone unless they were framed as ethical sensitivity. We used the resulting list to conduct citation chaining, searching reference sections for the keywords "sensitivity," "perception," "recognition" and "awareness." This process rendered a list of 336 papers.

As we read through these articles, we noticed some papers we judged to be of low quality (e.g., papers that included conclusions in the abstract that were not addressed in the methods or results and papers published in venues included in Beall's list of predatory journals [3]). We also discovered clusters of papers in regional journals that primarily cited other papers from their geographic region but were not well-connected to the international literature. We developed the following additional criteria for inclusion: papers must be peer-reviewed (excluding notes, dissertations, and most book chapters), must be from a journal with international readership and scope, and must be in a journal indexed by Scopus with a CiteScore of 1 or higher. On the first pass, 208 papers were removed: 36 dissertations, 69 papers in regional venues, and 103 whose publisher did not meet our CiteScore requirement. We confirmed that these criteria filtered out papers in predatory journals. Although these criteria may have unintentionally eliminated some high-quality papers, we are confident that our corpus represents the range of methods and indicators

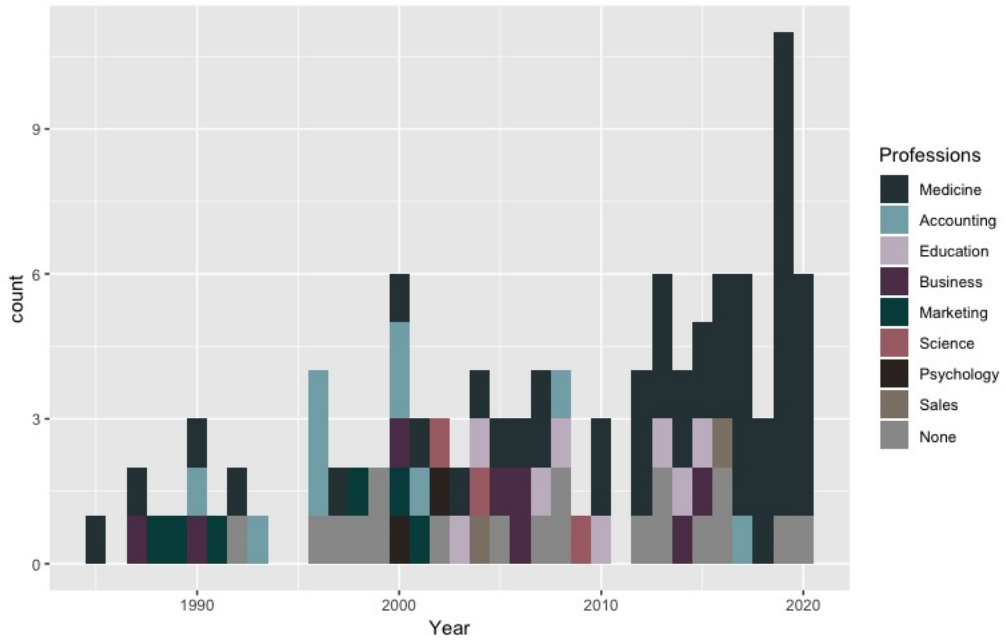


Figure 2.2: Papers in corpus by profession over time

used in current ethical sensitivity work. As we read through the remaining papers, we removed several more review, methods, and theory articles; the final corpus was comprised of 108 papers published between 1985 and early 2020. We read each paper and coded it for profession, jobs, sample size, methods, stimuli, measures, data type, whether the paper studied recognition, particularization, and/or judgment, and its conclusions. Details and examples of how we deployed these criteria are provided in Table 2.1.

Our final corpus reflected studies of wide variety of professions including medicine, psychology, accounting, education, law, journalism, and marketing (see Figure 2.2). These papers were published between 1985 and 2020 and studied between two and 1,971 participants. Research was conducted in the United States, Korea, Japan, Norway, Brazil, Turkey, China, and Taiwan among nurses, doctors, dentists, counselors, physiotherapists, managers, marketers, market researchers, insurance salespeople, accountants, teachers and students.

Table 2.1: Data Collection

Column	**Description**
Profession	What industry does this study address? i.e. “Medicine” or “Accounting.” “Business” is used when the paper uses it (i.e. studying MBAs). “None” is used when the study attempts to generalize to all workers or to people outside of a work context
Jobs	Specific jobs the paper is attempting to generalize to. i.e. “Nurses” or “Doctors.” “Students” is only used if the paper attempts to generalize to students within the profession “Education.” Papers using students to generalize about the profession they are studying (i.e. nursing students) the job are labelled (Nurses) and the use of students is recorded.
Sample	Number of participants in the study (range: 2 to 1,971, mean: 246, median: 165)
Method	Category of methods employed. i.e. “interview” “experiment,” or “survey”
Stimulus	The item participants will respond to. i.e. “scenarios” “survey items” or “reflection questions”
Measure	The name or citation for the measurement of ethical sensitivity, including citations not included in the corpus. If the paper uses a unique method, its own citation is used in this field. i.e. “REST,” or “Lind, 1993.”
Data Type	The format of data collected. i.e. “free response,” “brain imaging,” or “Likert”
Recognition	a. Does the paper’s measure of ES capture any aspect of recognition? b. A description of how elements of recognition are measured.
Particularization	a. Does the paper’s measure of ES capture particularization? b. A description of how elements of recognition are measured.
Judgment	a. Does the paper’s measure of ES capture particularization? b. A description of how elements of recognition are measured.
Conclusions	Summary of the study’s results

2.3 Conceptualizing ES

Weaver & Mitcham identified "gaps and discrepancies" in the literature about how to operationalize and conceptualize ethical sensitivity [185]. Other review articles and empirical articles join their call for better construct clarity [149]. ES researchers point out the difficulty of comparing results because of discrepancies among measures (e.g., [9]) and argue for more comprehensive or nuanced measurements [113, 137].

The literature reflects a lot of agreement on a single construct of ES: recognition. The moment of noticing an ethical issue is a dominant part of ES studies and separates it from studies of ethical decision-making and belief. But both empirical studies and review articles add varying additional constructs to the scope of ES. Sparks & Hunt [168] debate whether ES should be limited to recognition alone or include an assessment of how participants assess the importance of an ethical issue. Shaub [159] includes the identification of alternative actions and their outcomes, the awareness of consequences, and the actor's role in the situation. Jordan [106] notes that empathy and perspective taking may be worth including in ES. Weaver et al. [186] conducted a thematic analysis of the ES literature and identify eight attributes of ethical sensitivity. Concepts within recognition include affectivity, moral perception, and awakening. They also identify concepts beyond recognition in the literature to include particularizing, dividing loyalties, interpreting, justifying, and reflexivity.

No consensus exists on whether and what activities beyond recognition should be included in ethical sensitivity. For the purposes of this review, then, we separate recognition as a phenomenon of particular interest, and borrow Weaver et al.'s "particularization" to describe all activities that contribute to an understanding of the specific attributes of the situation at hand [186]. Our coding of the empirical ES literature suggests that particularization includes both information seeking and individual and group reflection. Information

seeking might include consulting professional standards [159], researching and stakeholders and consequences [113], and eliciting the perspectives of others (whether team members or external stakeholders) [153]. Reflection might include grappling with legal, social, or technical characteristics of the circumstance [186], considering the feelings and perspectives of others [153], and weighing personal attitudes and principles [120]). Sixty-four papers in the corpus included some aspect of particularization

Rest's original conception of moral sensitivity (published in 1982) was as the first of four discrete components of moral development: moral sensitivity, moral judgment, moral motivation, and moral character [148]. In later work (1999), Rest et al. acknowledged that sensitivity and judgment are intertwined:"Logically, Component 1 (sensitivity) precedes Component 2 (judgment), but the components do not follow each other in a set temporal order—as there are complex feed-forward and feed-backward loops, and complex interactions" [147]. While many conceptualizations of ethical sensitivity include recognition alone (e.g., [159]); recognition and some aspects of understanding [168]; or recognition and particularization [187]; some explicitly include judgment. For example, Byrd [36] (p. 14) defines ES this way: "Ethical Sensitivity is theoretically defined as the ability for one to recognize a situation as an ethical dilemma and to choose the action that is considered appropriate". Of the 108 papers in the corpus, 40 included a measure of judgment and eight measured only judgment.

Because social computing ethics is particularly interested in connecting the beliefs and practices of workers with the features of the technology they build, we believe it is useful to the field to include judgment as part of the scope of ES. However, because our literature review included only papers that framed themselves as being about ethical sensitivity, we excluded a large body of papers studying judgment alone (e.g., those framed as studying ethical decision-making.) Future research on ethical sensitivity in technology design may want to incorporate the ethical decision-making literature for more options about how to

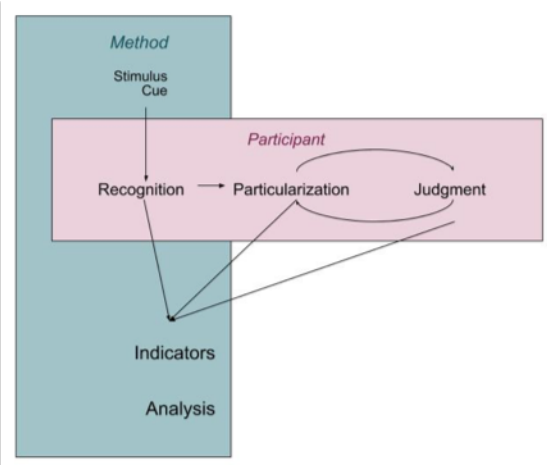


Figure 2.3: Operationalization decisions in ES studies

operationalize ethical judgment.

2.4 Methods and Indicators for Ethical Sensitivity

Even with the many component parts of ES condensed into umbrella concepts, our literature review revealed significant challenges in operationalization of those components: a common problem when studying unobservable theoretical constructs [100]. To operationalize and measure ES and its components, researchers must decide on a nested set of methods (what will be observed and how), stimuli (the situation(s) a participant will respond to), and cues (prompts for ethical reflection within the stimuli). Researchers must also decide upon indicators—observable properties of ES—and the analysis of those indicators. The relationships among these decisions are illustrated in Figure 2.3.

Analyzing the methods and indicators used in the ES literature revealed that methods for studying each component of ES— recognition, particularization, and judgment— are dominated by surveys in which participants were asked to respond to scenarios. However, the indicators used to measure or represent ES varied significantly based on the component(s) of ES a researcher was studying. Research capturing recognition was dominated by

survey responses in which the number of moral dilemmas respondents identified (primarily in vignettes) served as the primary indicator of ES. The indicators for studies of particularization were much more varied because of the complexity of building an understanding of a situated ethical issue. Studies of judgment were dominated by asking participants to evaluate the actions of others or suggest actions to be taken. We discuss methods and indicators for each stage of ES in more detail in the following sections.

2.4.1 Recognition

Recognition is the triggering mechanism of the ethical decision-making process: the moment a person identifies an ethical issue in their work [194][1]. While executing the technical tasks of their job (helping a patient, reviewing tax documents, or training a machine learning model, for example), a worker may perceive a signal—called a cue—that the situation requires ethical judgment. Recognition describes a perspective shift from seeing the task as primarily technical to ethical [51].

Much of the ES research defines and operationalizes ethical sensitivity as recognition alone, perhaps because whether a person identifies a situation as having ethical import is a feature that distinguishes "sensitivity" constructs from other ethical behavior frameworks like ethical decision-making [138, 176]. In our corpus, 56 papers focused on some aspect of recognition (51.9%). Figure 2.4 shows the methods and stimuli used to study recognition and Table 2.2 describes indicators of recognition in the corpus.

The ES literature revealed rather straightforward agreement on both methods and indicators for studying recognition. A majority of papers that measured recognition in our corpus used survey methods and employed scenarios as their stimuli (64.3%): vignettes of varying length, usually presented in writing, but sometimes video or audio, usually tailored

[1] Recognition is also sometimes called "awakening" or "perception."

Table 2.2: Recognition Indicators

Indicators	**#**	**Examples**
Naming moral dilemmas in a case	34	"Students were asked to list all the ethical issues related to the vignette they received. . . . scored by one investigator as to the number of issues the student identified, regardless of the content of the response" [88]
Is x an ethical issue?	10	"Recognition of an ethical issue (REC) was evaluated with one item indicating the extent to which the scenario involved an ethical issue" [92]
Proxy or proxies assumed to relate to ES	9	Image slides were shown in pairs, some depicting moral situations and some not. They were later revisited with some new slides included. "If participants more often correctly identify the novel person with morally-relevant interaction pairs [of image slides], it suggests superior encoding of morally-relevant stimuli. The same is true for social interaction pairs. If there is no increased recall of morally-relevant interactions or social interaction, it suggests that the brain is not wired to give more attention to moral interaction over social or non-social people perception." [133]
Brain scans	4	Participants were asked to look at images of in- and outgroup members harming or interacting peacefully with each other (4 conditions). "Each scenario was followed by 6 items designed to assess the participants' feelings of moral sensitivity toward the victim." Brain images & answers were compared [128]
How quickly participant mentions ethical issues	4	ES score weighted by time: how soon in their response participants mentioned each issue [116].
Rating importance of given factors	3	Participants were asked to rank considerations in decision-making. ES indicator was how highly participants ranked items belonging to the ethical perspective compared to organizational, personal, and legal perspectives [190]
Thematic Coding	2	"The numbered transcripts . . . were analyzed with inductive thematic analysis. After data immersion, two researchers respectively coded the transcripts to identify the main themes" including whether interviewees were sensitive to ethical issues in their job [95]
Self-report	2	The interview questions were the following: "Please explain your experience and perception of sensitivity in decision-making"; "How does your sensitivity affect your decision-making"; "Please explain what is meant by a manager's sensitivity in decision-making and what it includes." [151]

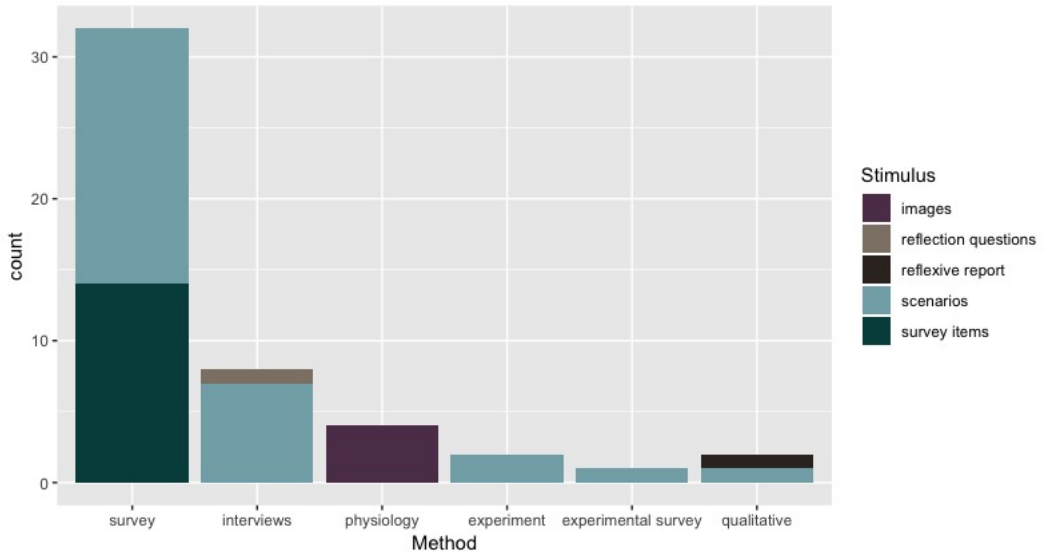


Figure 2.4: Recognition by Method and Stimulus

to the profession being targeted. Throughout the corpus, scenarios were most often employed in surveys (65.2%) but were also used in interview studies (19.6%), experiments (6.5%), experimental surveys (4.3%), as video prompts (2.2%), and qualitative studies (2.2%) The indicator of ethical sensitivity in scenario studies was generally how many ethical issues or how quickly the participant identified one. Alternative, less-frequent scenario approaches asked participants to indicate whether they believed issues or actions to be ethical and ranking factors (ethical or otherwise) used to make a decision.

After scenarios, a much smaller group of papers (21.4%) used survey items, which usually functioned like less-detailed vignettes. Recognition survey items frequently take the form of presenting a general behavior without backstory ("a nurse gives a patient his daily medication" or "a teacher ignores evidence of cheating") and the respondent is asked whether the action has ethical implications to observe whether the respondent recognizes the issue as ethical in nature. Rarer modes of studying recognition included pairing images with memory tests [133] or brain imaging [150] and free reflection on work experience, in which thematic coding by researchers identifies ethical recognition [90].

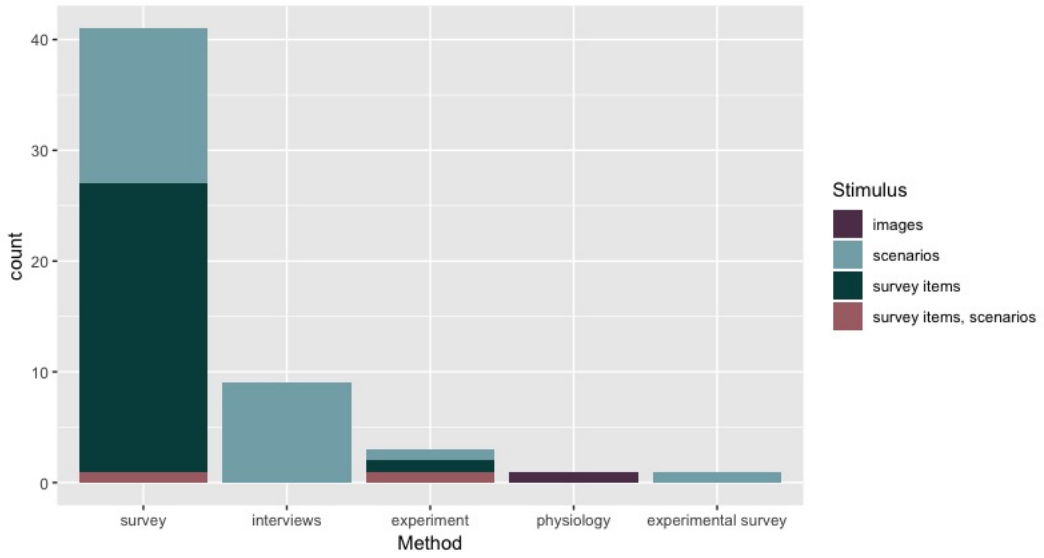


Figure 2.5: Particularization by Method and Stimulus

2.4.2 Particularization

After workers are alerted to the presence of a potential ethical issue by cues, they begin to particularize: to assess the scope and magnitude of the problem; determine its relationship to their own responsibilities; and understand the options and resources available to address the issue. Particularization "bridges the gap between moral rules (and principles) and particular situations" [24]. Particularization can be influenced by features of the situation (e.g., how much time and energy is available to process ethical cues), the cues (e.g., how many are available and what information they convey), and the individual (e.g., including personal values, individual reasoning, and social and embodied cognition) [186]. Workers also bring in external information during this stage, such as past experiences, the perspectives of other stakeholders, and external standards [186].

Sixty-four papers in our corpus studied particularization (59.3%). As was the case for recognition, surveys dominated the methods for studying particularization. When it came to indicators used to study particularization, papers were split between those that

Table 2.3: Particularization Indicators

Indicators	**#**[2]	**Examples**
Measuring theory-defined components of particularization	41	Used cognitive mapping to represent ES indicators mentioned by participants "in each of the four content domains (situational characteristics, issues, stakeholders, and consequences) and the linkages a subject makes between indicators" [113, 115, 116, 172]
Ranking or rating decision factors or ethical issues	10	Asked participants to rate the importance of each of several factors for each scenario: 'confidentiality', 'recipient of benefit', 'independence', 'seriousness of breach', 'recipient of responsibility', and 'growth of firm' [46].
Guided particularization	8	Walked participants in the treatment condition through an "Ethical Conflict Resolution Process" framework [123]
Evaluating thoroughness of response (range or depth)	7	In free responses to questions about a news clip, ethical sensitivity was measured by the range of different aspects of ethical sensitivity exhibited during the discussion (broader range taken to indicate higher sensitivity); and amount of thought, detail, sophistication (more depth taken to indicate higher sensitivity) [113, 115, 116, 172].
Labeling qualities of decisions	6	Likert questions in response to scenarios include: "Acceptable/Unacceptable" by tradition, family, people I admire, etc. "Efficient/inefficient," "in the best interests of the company/not" [31, 143–145].
Listing factors used to make a decision	5	"Following [brain] scanning, participants completed an open-ended questionnaire in which they were asked to indicate what types of information and strategies they used in determining violation severity ratings." [86]
What other information would you want?	1	"[Respondents] were asked to indicate the important . . . counseling aspects or issues related to the case, provide additional information they would like to know, and to suggest actions to be taken" [62].
Emergent codes	1	"The numbered transcripts . . . were analyzed with inductive thematic analysis. After data immersion, two researchers respectively coded the transcripts to identify the main themes" [95].

prompted respondents to focus on some forms of particularization (e.g., personal beliefs) and more naturalistic, reflection-based prompts. For example, the multidimensional ethics scale (MES), used in 4 papers in the corpus, guided respondents by asking whether an action taken in a scenario is acceptable or unacceptable to their family, tradition, or people they admire, emphasizing personal backgrounds [31, 143–145]. More natural particularization capture was prompted by open-ended reflection questions such as "What would you do to make yourself feel that you had done the right thing?" [61], asking participants what other information they would like to know in a situation [62], or asking participants to list any "issues," "concerns," or "aspects" of the case they'd use to make a decision. The Dental Ethical Sensitivity Test asked participants to carry on dialog from the scenarios as if they were in it, then asked why they responded this way, how they expected the patient to react, what issues in the situation are, what arguments could be made against their position, what the best interests of the patients are, what a dentist should do in this situation, and practically speaking, what they would do [19]. This series of questions was designed to "[require] the student to articulate the assumptions and perspectives underlying his or her response," using participants' identification of assumptions or perspectives as indicators of particularization [19], p. 227.

2.4.3 Judgment

After a worker has recognized and particularized an ethical situation in their work, they often go on to make a *judgment*. Rest's foundational work discusses judgment in three phases: formulating the morally ideal course of action; deciding what one actually intends to do; and executing and implementing what one intends to do [148].

Unlike other review studies (i.e., [106]), we found that judgment was studied in the fewest papers in our corpus: just 40 (or 37%). This is likely because we scoped our liter-

Table 2.4: Judgment Indicators

Indicators	**#**	**Examples**
Evaluate actions of others or an action in general	27	"Respondents were asked to indicate their approval/disapproval of the action(s) of the Marketing Research Director in each item (scenario). A 5-point scale with descriptive anchors that ranged from "disapprove" (coded 1), "disapprove somewhat" (coded 2), "neither approve nor disapprove" (coded 3), "approve somewhat" (coded 4) to "approve" (coded 5) was used to elicit their evaluations" [10].
Suggest actions to be taken	11	"Practically speaking, what would you do?" [19, 130].
Choice of actions	6	"Subjects responded to a total of sixteen ethically sensitive situations of a personal or business nature. For each situation, subjects were asked to indicate their probable action on the issue on the five point scale of yes, probably yes, unsure, probably no, and no" [139].
Justification	2	"Participants were encouraged to articulate an argument and rationale describing whether they would support the use of [the therapy described in the scenario]" [153].
Assign a penalty	2	". . . asked the participants' opinion regarding the penalty that they thought would be appropriate had there been a higher council that ruled the action in the vignette as unethical. Five potential penalties were presented in order of most severe alternative (i.e., the person engaged in the activity should be taken to court), to no penalty (i.e., there is no need for punishment). . .participants were asked to indicate their degree of agreement with each penalty suggestion using a seven-point Likert scale" [167].
Evaluate the character of others	1	Participants rated characters on "their perceived morality on a 9-point scale (1: very immoral character, 5: neither moral nor immoral, 9: very moral character) [117].

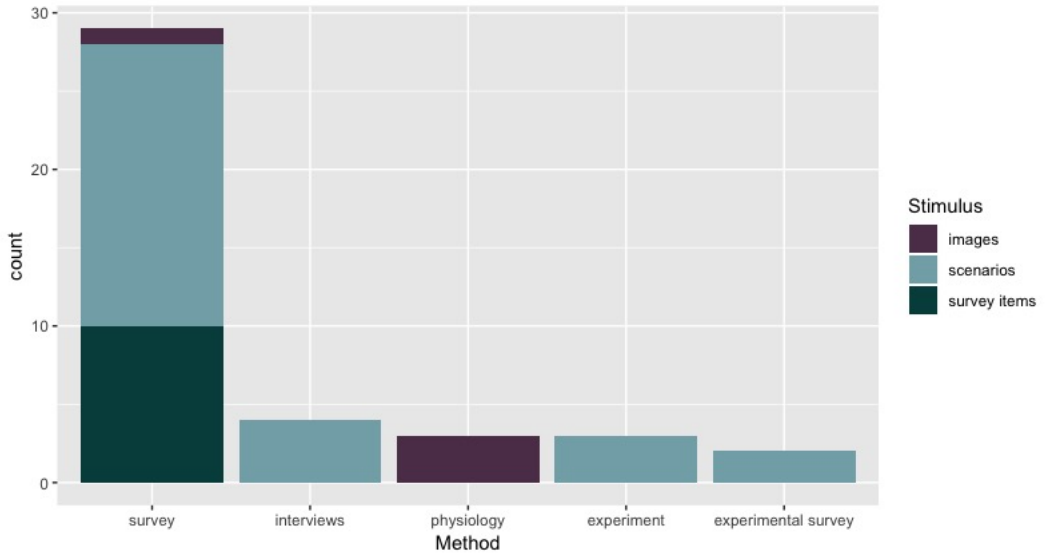


Figure 2.6: Judgment by Method and Stimulus

ature search around ethical sensitivity-specific language and excluded papers that studied decision-making without making reference to the larger scope of ES. [3]

Like recognition, judgment was studied most often through surveys, using scenarios as stimuli (45%). Figure 2.6 reviews the methods and stimuli used for judgment in this corpus. Indicators of judgment included participants' evaluations of the decisions of the characters in a scenario, participant's responses to whether they would perform a particular action, participants' opinions about what should be done, or participants' explanations what they would do if they were a particular character in the scenario (See: Table 2.4).

[3] Pederson conceptualized the relationship between ethical sensitivity and ethical decision-making, linking ethical sensitivity to problem formulation, the first step of some ethical decision-making frameworks [138]. He also argues that experience with ethical decision-making feeds back into the development of ethical sensitivity. So, these two concepts are quite related, and while we've separated them for expediency here, researchers interested in ethical judgment may be interested in the broader literature on ethical decision-making.

2.4.4 Studying all Components of ES

Most studies in the corpus (62 papers, or 57.4%) focused on a single aspect of ethical sensitivity, studying either recognition, particularization, or judgment. Sixteen studies, however, focused on all three components (14.8%). Of the 16 papers that addressed recognition, particularization, and judgment, nearly all (14) used scenarios as the stimulus and used coding of participants' free responses as indicators of recognition, particularization, and judgment (12). Most of these responses were gathered via survey (9), with a handful gathered via interviews (3).

2.5 Studying ES in Social Computing

How might we draw on methods, indicators, and research experience from the existing ES literature to conceptualize new research in technology ethics? And what does the HCI and CSCW literature add to the study of ES? The ES literature benefits CSCW and HCI research by adding conceptual specify, shared language, and validated indicators to studies of the ethical practices of technology designers. Explicitly observing recognition (what cues designers to notice potential ethical issues); particularization (what information sources, norms, stories, and beliefs they draw on to build their understanding of issues, options, and consequences); and how both recognition and particularization shape design and policy decisions gives technology ethics researchers a shared set of constructs to observe, helping us both design and present our research and compare findings across studies. Conceptual specificity and validated indicators will also improve the design and evaluation of the many ethical interventions developed by HCI and CSCW researchers. This section discusses how recognition, particularization and judgment can scaffold studies of, and interventions into, tech ethics work.

Simultaneously, HCI and CSCW researchers add creativity in methods and stimuli choices for studying ES in design work. Studies of ethics in technology design have employed more diverse methods, including ethnography, document analysis, and quasi-experimental methods to study dimensions missed by the current ES literature: in particular, the interactions of teams, and the situated and embodied nature of ethical sensitivity. In the following sections, we link the components of ES to existing HCI and CSCW research to illustrate how ES can benefit technology design research, and how technology design research can expand ES.

2.5.1 Recognition in Social Computing

Explicit attention to ethical recognition in technology design can help us iterate on and improve the many ethics interventions developed in the HCI and CSCW literature. Ethics interventions such as design workbooks and card sets function as cues for ethical recognition. Conceptualizing interventions as cues for recognition enables us to ask evaluative research questions such as: do some interventions increase ethical recognition more than others? The ES literature suggests indicators to quantify the impacts of interventions, such as asking participants to name the moral dilemmas they encountered after an intervention, asking if participants recognize particular ethical issues, or measuring how quickly participants can name ethical issues.

The ES literature can also scaffold HCI and CSCW literature that seeks to discover naturalistic cues for ethical reflection. For example, the framework of ethical sensitivity helps us see values levers [160, 162, 163] as one type of cue for ethical recognition, and ES can help discover additional cues. Building upon ES studies such as those conducted by Wittmer et al. [189, 190], which asked participants to engage in realistic managerial decision-making exercises, design researchers might ask participants to work through a

design activity or reflect on their previous work experience to identify previously undiscovered workplace cues for ethical reflection [11, 90]. Using more established ES methods like surveys and scenarios can also facilitate new discovery in CSCW and HCI. Researchers might draw from the validated scales used in the Dental Ethical Sensitivity Test [19] and Racial Ethical Sensitivity Test [26], which use open-ended questions to elicit recognition and understanding of specific ethical issues relevant to context (dentistry and education, respectively). Researchers can use a "funnel sequence interview" such as that developed by Lind and Rarick, using strategically ordered questions about ethical issues that progress from less to more direct [114]. The funnel sequence interview is used to obscure the purpose of the interview [114] and to give participants a chance to signal recognition of ethical issues before being prompted to do so [113].

CSCW and HCI studies of technology design can offer a richness of data and context that is missing from existing ES studies of recognition. Analyzing ethical recognition in secondary or historical data, for example, is a method that has been employed to study ethical reflection by HCI researchers but was not observed in the ES corpus. Email archives and online discussion forums can offer a view into individual and group moments of ethics discovery, including among workers distributed over time and distance [164]. Connecting contextual methods to the language of recognition and cues offers HCI and CSCW researchers a way to both define indicators for future comparative studies, and to connect moments of noticing (and factors that impact those moments) to design-critical next steps such as ethical particularization and judgment.

2.5.2 Particularization in Social Computing

Particularization is perhaps the richest opportunity for CSCW and HCI researchers to contribute data and methods to ethical sensitivity research, and to improve our understand-

ing of design ethics by connect existing work focused on noticing ethical issues to design and policy decision-making. Particularization during technology work is a complex phenomenon to study because it can be influenced by a host of both personal or group (e.g., education, experience, beliefs, communication, and practices) and situational factors (e.g., organizational context, strength of the cues, time available to process cues, pressures from other priorities). The ES literature has defined a numerous ways of measuring components of particularization. Many of these are purposefully reflective, such as asking participants to list and rank the factors that went into a decision and evaluating the range and depth of those responses. These reflective techniques can help CSCW and HCI researchers add new concepts for observation, measurement and evaluation to studies of ethical practices in technology work.

Established features of particularization that could advance HCI and CSCW studies include assessments of ethical issues priority among teams and leadership as in [159], listing factors influencing design decisions as in [35], evaluating designers' perception of stakeholder interests and conflicts as in [42, 113], surveying designers' acknowledgment of moral ambiguity as in [42], measuring designers' awareness of alternative courses of action available as in [42], assessing designers' understandings of how other actors might respond to alternative courses of action as in [26], studying designers' perceptions of consequences of courses of action as in [113], evaluating whether designers anticipate arguments against a course of action as in [26], and studying whether and when designers refer to external information sources as in [186].

Another particularization concept defined in the ES literature that could be useful to CSCW and HCI researchers is developing moral imagination as the skills useful to, and necessary for, particularization [104]. Imagination is required for empathy with other stakeholders, to search for relevant concepts and principles, and to consider all available options [47]. And while there is a broad range of work on improving moral imagination

in computing (e.g., pedagogical techniques like case studies and role playing [66], studies of moral computing exemplars [96], and design fiction activities [63, 192]), these methods are rarely evaluated for their impacts on particularization. The ES literature suggests next steps for studies of moral imagination such as studying the effects of organizational socialization [168], job satisfaction [194], and other conditions of work on moral imagination, especially in teams. Ongoing research into particularization processes in technology ethics could improve not just research on technology work, but research on ethics pedagogy. Social computing research can also contribute rich insight into the ways that situational factors impact particularization in technical practice. Although existing ES review papers describe particularization as a process of building understanding [24, 186], a majority of empirical papers in the ES corpus focused on capturing the post-hoc ethical understanding participants have of a situation. Social computing methods such as talk-aloud studies [60], trace ethnographies of ethical decisions [75], and ethnographic observation of values reflection among developers or users (e.g., [13, 94]) offer opportunities to observe the complex interactions between personal and situational factors during the process of particularization. For example, Gray led students through a design process that asked students to generate an "elevator-pitch" for their design concept [78]. He observed that students had difficulty with "ethical awareness" and was able to pinpoint challenges building empathy with users, which is a step that could be considered a component of particularization. In other work, Gray and Chivukula used on-site observations and follow-up interviews to observe ethical decision-making in the context of organizations [79]. They recorded work practices, speech acts, and non-verbal signals during the observation period, and documented participants' reflections during the interviews. This method fills a gap in ethical sensitivity research by focusing on in-situ ethical practices.

Finally, several aspects of particularization that are described in ES theory, but underexplored in the current ES literature, may be of particular interest to social computing. In-

terpretation, or contextual translation, is a particularization activity during which a worker reflects critically upon the influence of larger social systems by acknowledging their own assumptions or privilege and explicitly considering the perspectives of other internal and external stakeholders [186]. Numerous CSCW and HCI ethics interventions (e.g., envisioning cards) ask developers to reflect on the intersection of systemic factors and privilege in their work. Envisioning cards could be used to test how this form of particularization impacts teams' design decisions. Dividing loyalties describes how workers resolve tensions among conflicting values and interests of multiple stakeholders [187]. In technology work, methods for grappling with divided loyalties have been formalized as stakeholder analysis techniques (e.g., values dams and flows [126]) in values in design and value sensitive design literature [162]. Reframing these methods as an explicit part of ethical particularization again allows us to connect the impact of methods like values dams and flows to design decisions in real-world settings.

2.5.3 Judgment in Social Computing

Because social computing and design ethics is interested in connecting the process of perceiving ethical issues to design outcomes, studying judgment – and the relationship of judgment to design – is a final important component of ES for our field. Our literature review revealed that studying judgments is frequently easier to operationalize than recognition or particularization, both because people may recall their decisions with more precision, and because these decisions can often be seen in documentation (whether written policies or design decisions). The relative ease and importance of studying judgment extends to current work in technology ethics and social computing, as well. Research such as Fleischmann et al.'s work [67], which asks engineers to reflect on ethical decisions using both surveys and interviews, or the wide body of literature that examine technological

objects for purposeful and implicit values within their design [134, 179], are examples of operationalizations of judgment in tech ethics research. Content analysis of design decisions, such as those by Shilton and Greene [164], or of user reactions to design decisions, such as that by Gray, Chivukula, and Lee [77], suggest document-based methods for studying judgment. And observational or ethnographic work focused on design decisions, such as those used by Chivukula et al. can trace the full path from stimuli to cues to recognition, particularization, and judgments [43]. We believe that explicit reference to the ES literature can improve and contextualize work focused on judgment by providing new research questions into the factors (cues, recognition, and the many components of particularization) that impact judgments. ES gives us concrete language to ask: what factors impact ethical design decisions?

2.5.4 Studying Ethical Sensitivity as an Aspiration

While we have largely discussed ethical sensitivity as a descriptive framework for studying design work, the construct also has normative potential. For example, existing research suggests that ethical sensitivity can be learned through both professional education and socialization within a field [184]. However, there is no consistent ethics curriculum among computer science departments [64], nor do all developers receive a computer science degree [6]. Although developers can belong to a professional organization with a code of conduct (e.g., Association for Computing Machinery or the Institute of Electrical and Electronics Engineers), membership is not required to practice. Professions with strict occupational closure require their workers to complete educational programs and maintain credentials, which offer a way to consistently discuss ideas, disseminate best practices, and enforce rules. Software development instead relies on individual workers to notice ethical issues and determine what to do. Reframing ethical sensitivity as a normative goal enables

us to think through indicators for evaluating computer ethics education and the impact of professional ethics efforts. For example, explicit goals of computer education might be increasing recognition of workplace ethical cues, or increasing the forms and modes of particularization students to which students are exposed.

Similarly, Weaver & Morse suggest that ethical sensitivity can fail to develop (e.g. workers miss cues, or do not engage in particularization and judgment) if workers either are certain about the right thing to do in a given situation, or if they address situations as solely technical [187]. Evidence from previous studies of ethics in software development suggests that this may be a common situation in development work. For example, Shilton has studied how important technical values like interoperability are frequently translated as values neutrality, interfering with ethical reflection during design [160]. Cech has found that engineering students' engagement with questions of social welfare issues declines over the course of their college engineering education [40]. To counter these tendencies, Lurie & Mark have proposed a form of software development that explicitly interweaves ethical frameworks into technical decisions to address a perceived separation of technical work from stand-alone codes of ethics [119]. The IEEE Standards Association has proposed P7000, a process for ensuring ethical concerns are addressed during system design [5]. Building explicit reference to recognition, particularization, and judgment components of ethical sensitivity into ethical design frameworks such as these could help the field form shared vocabulary around the steps that ethical reflection requires.

Finally, we can also, like Heggestad et al., use ethical sensitivity to reflect on our own research practices [89]. Considering how HCI and CSCW researchers recognize ethical issues in their work, particularize those issues with their own beliefs, the practices of their colleagues, and reference to external guidance ranging from ethics review boards to professional code of ethics, and then make judgments about ethical research practices, could further the fields' growing interest in the ethics of its own research practices [27–

29, 65, 129, 181, 196].

2.6 Conclusion

This paper is the result of two HCI researchers discovering, analyzing, and ultimately critiquing the literature on professional ethical sensitivity. The lead author, a doctoral candidate, was seeking a theoretical framework that could guide her investigations of when and how machine learning engineers consider and address issues of bias in their work with training data. Her discovery of a broad ethical sensitivity literature was exciting: it provided both conceptual foci (where to look) and suggested methods (how to look) for ethics in practice. The second author, her advisor, instantly recognized components of her own work within the ES framework. She realized she had long been studying a component of ES – recognition of ethical issues in technical work, and the cues that trigger that recognition – without naming it as such. And she recognized that numerous other HCI scholars had been studying another component of ES – judgment – in the form of design decisions. We were inspired to systematically review the ES literature to understand how scholars outside of social computing have used these terms, and how they have observed or measured these constructs. This paper traces our journey through the Ethical Sensitivity literature. We remain excited for the concrete and well-justified framework – the what to study – that ethical sensitivity provides to studies of design workplaces and social computing. But we found less than we'd hoped for in guidance of how to study phenomena such as ethical recognition, particularization, and judgment in ways that take into account cooperative work and situated and embodied practices. The survey methods that dominate the current ES literature is quite different than the methods that dominate CSCW and HCI ethics research, which tends to observe participants' situated ethical sensitivity by studying teams [80], by observing real-life work situations [165], or by providing real-world stimuli [192].

Despite the gulf between the constructs measured in the ES literature and those observed and interpreted in CSCW and HCI, our analysis does suggest ways that a broad range of researchers interested in ethical sensitivity in technology work can structure their research. First, our analysis distills from the literature three concepts particularly useful for studying cooperative design work: a team's recognition of ethical issues, particularization of those issues through both internal reflection and external information seeking, and judgment through actions, policy, or design decisions. A wide variety of indicators have been explored and validated in the current ES literature, particularly for recognition (which usually involves counting or classifying the ethical issues named by participants) and judgment (which focus on action, either counting the actions recommended by participants or asking participants to evaluate the actions of others). Adapting these for studies of technology design might include counting the first mentions of ethical issues among development teams, classifying the judgments and justifications offered by workers, and asking participants to connect the options and reasons they considered to the actions they executed.

The technology studies literature is increasingly focused on changing the conditions of technology work [81, 125], recognizing that even the most ethical individuals cannot create change unless the entire sociotechnical system allows for that change. ES is a useful framework for understanding ethics in work practice and workplaces, but significant new empirical research is needed to define the factors most important to particularization within tech workplaces, alongside the stimuli and cues that are most effective for recognition of ethical issues, and how these two components interact to result in ethical judgments, policy, and design decisions. As a student and a seasoned researcher, we still struggle to operationalize these fuzzy social constructs in our own work. But narrowing them from studying "ethics" to studying recognition, particularization, and judgment gives us useful signposts to apply our own embedded, observational methods. We hope other CSCW researchers will find similar clarity and utility in the ES literature.

Chapter 3: Datasheets for Datasets help ML engineers notice and understand ethical issues in unfamiliar training data

Abstract The social computing community has demonstrated interest in the ethical issues sometimes produced by machine learning (ML) models: violations of values like privacy, fairness, and accountability. This chapter discovers what kinds of ethical considerations machine learning engineers recognize, how they build understanding, and what decisions they make when working with a real-world dataset. In particular, it illustrates ways in which Datasheets for Datasets, an accountability intervention designed to help engineers explore unfamiliar training data, scaffolds the process of issue discovery, understanding, and ethical decision-making. Participants were ethically sensitive enough to identify ethical issues in the dataset; participants who had a Datasheet did open and refer to it; and those with Datasheets mentioned ethical issues during the think-aloud earlier and more often than than those without. In addition to encouraging data about the use of Datasheets in particular, this study offers evidence for the promise of similar context documents and a means for testing interventions that claim to encourage recognition, promote understanding, and support decision-making among technologists.

3.1 Introduction

Machine Learning (ML) finds patterns in training data, but doesn't distinguish between useful bias (that helps it differentiate between images of cats and cars) and discriminatory

bias (that, for example, assesses Black parolees as more likely to reoffend [14]). It reflects bias that exists in the world from which the data was drawn, for example in associations between words in text, resulting in algorithms that reify those biases [37]. Training data has also become a target for accountability interventions [21, 44, 74, 91] and is described by industry practitioners as a key place to intervene to support fairness in ML [93]. Training data are also an area of particular concern for privacy advocates, but the issue is complicated because collecting sensitive attributes (e.g., race, gender, or indicators of class) may be necessary to build and certify fair algorithms [183]. Therefore, researchers advocate interventions into training data to preserve privacy and allow for fair model training, model certification, and decision verification by, for example, encrypting sensitive attributes in training data [108]. Because training data is cited as a driver of discrimination, a focus of concern for those concerned with multiple ethical issues, and highlighted by industry practitioners as a target for intervention, this project focuses on interventions that deal with data collection, manipulation, and training.

Context documents are designed to accompany a dataset or ML model, allowing builders to communicate with users. These documents ask dataset or model builders a variety of questions: some ask about the context of development or data collection, measures of data distribution or model performance, ethical and legal concerns, but most ask questions from more than one category. Many were proposed in part to prompt technologists to recognize and develop a thorough understanding of ethical issues [21, 74, 146, 158, 166, 193]. As part of that effort, most of those include direct ethical questions. For example, "Were any ethical review processes conducted?" "Are there any tasks for which the dataset should not be used?" and "Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups?" [74]. Others take another approach: in their

paper proposing Data Statements for Natural Language Processing, Bender & Friedman argue that their proposed context document may surface bias and other ethical problems without a question directly about ethics: "We propose here that foregrounding the characteristics of our datasets can help, by allowing reasoning about what the likely effects may be" [21]. This study offers some empirical data to support the idea that indirect questions about dataset characteristics can prompt ethical engagement.

Ethical sensitivity (ES) gives us a framework to observe how context documents scaffold this kind of ethical engagement. ES describes a moment of recognition (where someone working on a technical task notices its ethical aspects), particularization (where the worker seeks information and reflects to build an understanding of the specifics of the situation), and judgment (where the worker selects and executes a path forward). It has been studied in a wide variety of professions [186] and has been used to test educational interventions [45, 48, 102, 130], but has never been used in technology development nor been operationalized for in situ work data.

This chapter uses the construct of ethical sensitivity describe how a selected context document– Datasheets for Datasets [74]– influences recognition and whether and how it shapes particularization. To do this, I observed ML engineers after I presented them with an ethically problematic ML problem and data, some with a Datasheet and some without. I asked participants to think aloud as they worked with the data and watched them recognize and particularize

This study answers the following research questions:

> **RQ 1**: Are engineers who read Datasheets cued to perceive ethical problems differently or at different rates than those who do not read them?
>
> **RQ 2**: What information on and off the Datasheet do ML engineers use to particularize a perceived ethical problem?

More participants who were given Datasheets recognized ethical issues while working with the data and participants relied heavily on the Datasheet to particularize.

Section 3.2 reviews current work on ethical sensitivity and ethical topics with training data. Section 3.3 describes methods, 3.4 discusses analysis, 3.5 explains the results, and 3.6 reviews the importance of these findings for ML development, ethical sensitivity, and ethical cooperative development.

3.2 Literature review

I drew on several areas of existing work to develop this project. First, I will discuss the role of training data in ML and how fairness, privacy, and accountability in an ML system can be affected by training data. Next, I'll introduce the category of interventions I call "context documents" and explain how I selected one to focus on. Finally, I'll review the framework of Ethical Sensitivity and how context documents try to promote two of its components: ethical recognition and particularization.

3.2.1 Training data and ethics

Although there's some debate about what qualifies as machine learning, the defining feature is in the name: "learning." ML algorithms are said to learn patterns by automatically and iteratively optimizing a model to fit training data. For our purposes, it won't be necessary to precisely delineate machine learning from traditional software and statistical methods; for the purposes of this study, "machine learning" refers to algorithms that find patterns in training data and use those patterns to classify, predict, or do some other task without being explicitly programmed with rules for doing so.

There are several human values of interest relevant when considering the training data used to build ML models. Fairness, privacy, and accountability are of particular interest to

the facial recognition dataset used in this study.

Fairness

Training data can be a table of rows with features and a dependent variable, like a regression would need, but training data can also be images or video, unstructured text, shopping histories, online learning activity, and much more.

Training data is biased. Most of the time, that bias is exactly what the algorithm uses to accurately label an image of a cat rather than image of a dog or a person or a car. But it can also contain bias that reflects real world prejudice or systemic discrimination, and an algorithm can't tell the difference between "a longer nose-looking thing tends to correlate with the label 'dog' and smaller, shallower face is more likely to be 'cat"' and "the phrase 'boy scouts' predicts an interview, but when a resume contains 'girl scouts' it tends to go in the 'no' pile." Examples of this outcome bias abound (For reviews, see: [17, 136])

Discrimination in ML models isn't only decisions reflecting societal prejudices, it can also manifest as performance differences between groups. A striking example of this is the "Gender Shades" paper, which found that facial recognition performed worse for darker-skinned subjects and female subjects, with error rates for darker-skinned women as high as 34.7%, when the highest error rate for lighter-skinned men was less than 1% [33].

Privacy

As ML pervades new domains, so does data collection. Targeted advertising, facial recognition, recommendation services, search engines, spam filters, and self-driving features in cars mean that browsing history; photos of people online, in public, and in public records; viewing, listening, and purchasing histories; email; and driving behavior are being collected. Many people have raised privacy concerns about this data [161] and about

"notice and consent," the ethical safeguard used for collecting much of this data [16].

Including many, diverse examples in a training data set can address quality problems, including performance bias, but adding more data can mean more people's privacy is at risk. This is aggravated by the fact that in order to identify discriminatory bias, sensitive attributes may need to be collected and stored [183]. Interventions to address bias may prescribe oversampling rare or sensitive cases, meaning that members of minority groups can be more likely to have data collected about them than those of majority groups.

Accountability

Many advocate for transparency into algorithmic decisions to allow its builders, users, citizens, courts, and regulators to understand, interpret, and act on their recommendations without creating ethics and policy violations. There is more than one avenue to this goal, including transparency– reporting on methods, data, and models– and interpretability– techniques that offer a view into the workings of models. This study focuses on a particular type transparency intervention I'll call "context documents."

3.2.2 Context Documents

Sometimes, the same team collects training data and prepares it to train a model, but not always. It may be different teams' responsibility in a large organization, engineers may reuse data collected for other purposes (for example sales, quality control, or user data), or they may use any of many large, public datasets available. OpenML lists more than 2,600 open datasets [178].

Recently, researchers have called for standard documentation accompanying datasets or ML-driven systems that describe characteristics of training data and help others interpret and use the data or model. These "context documents" function as a way to communicate

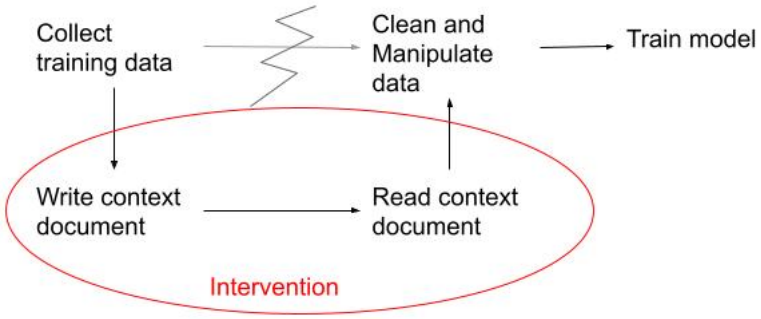


Figure 3.1: Context documents used to bridge the gap between data collectors and users

information about data provenance from those who collect it to those who use it. For example, Figure 3.1 shows how context documents for pre-processed data work to overcome a disconnect between those who collected the data and those who use it.

Context documents take many forms, ranging in complexity from a few hundred words [156] to detailed reports [21, 91]. Proposals like Bender and Friedman's [21] for Natural Language Processing and Yang et al.'s [193] for ranking algorithms illustrate the specificity that context documents tailored for a single ML technique can offer. Some are part of larger programs or regulatory regimes and have a format tailored to their purpose in it [146, 158, 166]. Gebru et al.'s Datasheets [74], Mitchell et al.'s Model Cards [127], and Yang et al.'s nutritional label [193] directly ask for information about ethical concerns, while others argue that simply reporting the characteristics of datasets will prompt and advertise ethical work [21, 91]. The sudden proliferation of context document proposals may be a response to an uptick of research and journalism verifying algorithmic bias: all but one of these context document proposals cited either Julia Angwin's "Machine Bias" [14], Bolukbasi et al.'s "Man is to Computer Programmer as Woman is to Homemaker?" [25], or both.

Context documents are designed to intervene not in the technical product, but in the practices of technology designers. Documents have been proposed for both pre-processed

training data and completed models (at the beginning and end of the training data resourcing cycle). Tables 3.1 and 3.2 classify each of the context documents mentioned above on this and some other key dimensions.

Table 3.1: Context Documents by Scope and Focus

	Technique- or Domain-specific	**General Purpose**
Training Data	Bender & Friedman, 2018 *(Natural Language Processing algorithms)*	Holland et al., 2018 **Gebru et al., 2018**
Model	Daikopoulos, 2016 *(Media & Journalism)* Selbst, 2018 *(Policing)* Yang et al., 2018 *(Ranking algorithms)*	Mitchell et al., 2019 Reisman et al., 2018 Shneiderman, 2016 Schmaltz, 2018

Table 3.2: Context Documents by Audience and Purpose

General Call/Programs	Technical Reports	**Lay Language Documents**
Shneiderman, 2016	Holland et al., 2018	**Gebru et al., 2018**
Reisman et al., 2018	Bender & Friedman, 2018	Bender & Friedman, 2018
Daikopoulos, 2016	Yang et al., 2018	Schmaltz, 2018
		Mitchell et al., 2019

This study focuses on Datasheets [74]: the only technique- and domain-agnostic, lay-language context document for training data. Datasheets are versatile: they can be taught early in ML education to students who will go on to work in diverse domains using a variety of techniques and can be read by non-experts, like managers, users, citizens, and auditors. Although Datasheets aim for a variety of goals [74], one of them is to increase the likelihood that ML engineers notice, understand, and can act on ethical problems in datasets.

3.2.3 Ethical sensitivity

Ethical sensitivity started as Rest's "moral sensitivity" [148] and has emerged as a way to understand how people recognize, interpret, and act on ethically consequential decisions in their work [186] (see Chapter 2).

Although ethical sensitivity is often studied as a trait (e.g., by asking "are men or women more ethically sensitive?" [10, 139]) some studies suggest that it may be a skill that can be developed [56], taught [68], or a collection of related skills [113].

This chapter treats ethical sensitivity as a practice, not a trait, for which a person or group can be more or less disposed, more or less skilled, but which is capable of being developed. In other words, in contrast to a survey that tries to measure ES as a latent trait (like the popular Moral Sensitivity Questionnaire [121]), this study treats ES like an ethics-focused application of Aristotle's *phronesis*, a practical wisdom that bridges *techne* (context-dependent knowledge of one's craft) and *epsiteme* (universal, unchanging knowledge). Phronesis has been profitably used to understand professional reflection and judgment [110].

According to this view of ES, a person may consciously practice ethical sensitivity at work, in a classroom, in a role-playing exercise, or other simulated scenario. A worker may also perform it as part of their work without consciously exercising their skill. It can be supported or undermined by organizations' practices, policies, and people, like job descriptions, evaluation schema, and managers. This view of ethical sensitivity strikes a chord with long-standing research models that assert that ethical decision-making is dependent on the worker and their work characteristics [175] and influenced by the particulars of the ethical issue [105].

ES is highly situated, but that doesn't mean we can't learn about it in a simulated work environment, just as skill at basketball can be developed and meaningfully observed outside

the context of a team, a game, an audience, a league, or even a particular set of rules by watching players practice. The practice session can be thoughtfully designed to develop (or demonstrate) skills that carry over to in situ performance, including focusing on areas of particular weakness or interest. Similarly, a curriculum, training activity, or study can develop or demonstrate ethical sensitivity.

For this study, it was important to observe people who had a Datasheet and those who didn't with the same problem statement and to be able to control the demographic distribution of the training data, so observing people in their full, long-term work context was not feasible. Therefore, I developed a think-aloud method that would allow ML engineers to participate from their work environment on their own machines with their preferred settings and software.

If we assume that we can observe ethical sensitivity in a study, what are we looking for? A variety of common features have been identified in studies across disciplines and will be described in the following sections.

Recognition

Recognition of an ethical issue is the moment that gives a worker the opportunity to intervene. While executing the tasks of their job (helping a patient, reviewing tax documents, or training an ML model, for example), a professional may perceive information that signals that the situation requires ethical judgment: a perspective shift from seeing the task as primarily technical to ethical [51, 139].

There's little prior work in ethical sensitivity describing or categorizing what Weaver et al. refer to as "cues" that trigger ethical recognition and guide the first steps of particularization [186]. Context documents like Datasheets may operate as cues, and in fact, some are designed to do so. Papers proposing these documents talk about their possibility for

allowing dataset or model users to "recognize . . . potential limitation" [21]. Holland et al. [91] argue that the Dataset Nutrition label they propose may highlight characteristics of data and enable engineers to "check for issues at the time of model development." Mitchell et al. [127] motivated their proposal in part by noting that some "systematic errors were only exposed after models were put into use," hoping that Model Cards could help avoid such oversights.

Particularization

Particularization is a less well-defined and -studied area of ethical sensitivity. Weaver et al. [186] include activities that develop an understanding of the particulars of the ethical situation, such as reflecting on one's own beliefs, seeking information about the circumstances, and referring to external standards, like policies or codes of ethics. Blum [24] explicates the importance of such particulars to ethical judgment. Chapter 2 reviews aspects of particularization that have been included in ES work. For the purposes of this study, particularization is defined as any kind of understanding-building activity.

Context documents are not only created to spark recognition. Bender and Friedman (2018) say that their report could "[allow] reasoning about what the likely effects may be." Mitchell et al. [127] discuss several targets of particularization as goals for their document, including how the cards can help stakeholders identify what questions to ask of a model and evaluate its suitability for a given context. Schmaltz [156] highlights the ability of a context document to cause builders to consider societal implications, risks, and failure cases. Datasheets were designed to help readers evaluate the appropriateness, strengths, and weaknesses of the dataset is for their purposes, and to encourage creative, critical thought on the part of the Datasheet authors about the dataset [74].

The process of building an understanding of the particulars of an ethically consequential

situation is under-explored in the existing ES literature (See: Chapter 2). The think- aloud method may offer insight into particularization that isn't available with the survey methods used extensively in that literature.

This study uses ethical sensitivity to evaluate the effectiveness of Datasheets, especially whether they help ML engineers recognize and particularize ethical issues in a dataset.

3.3 Methods

This study seeks to understand how introducing Datasheets may spark ethical perception, inform particularization, or otherwise change engineers' practices when exploring a new, ethically complicated ML problem and data set. First, I wanted to know, when given the option to do so, how many ML engineers read a Datasheet when faced with an unfamiliar data set and ML problem? Then, among those who read the Datasheets, what information do participants refer to while working on an ML problem, first noticing an ethical problem, or particularizing one when they see it? Do Datasheet readers recognize and engage differently with an embedded ethical issue than those who do not read a Datasheet?

To get this data, I asked 23 ML engineers to think aloud while exploring a data set and problem statement with multi-faceted ethical problems.

3.3.1 Participants

Participants were recruited through the Slack channel for an ML meetup group the author attends (6), referral from other participants (7), and several forums (/r/machinelearning, /r/artificial, /r/datascience, and hackernews.com). They were offered a $40 Amazon.com gift card for an hour session. Participants needed to be 18 years or older and consider themselves data scientists, machine learning engineers, or people who worked with training data data science or ML algorithms. Participants experience and job roles are described

Table 3.3: Participants

Experience	Min	.1
	Max	15
	Average	3.5
Job Role	Worker	12
	Student	8
	Manager	2
	Volunteer	1
Industry	Industry	15
	Academia	2
	Both	2
	Unclear/ Unsure	4

in Table 3.3. Three participants were primarily self-taught, and, in addition to university classes, other participants reported learning through online courses (participants mentioned Coursera (6), Udacity (2), and Stanford online (1) specifically). Several participants were in or had recently completed a mentored, self-paced bootcamp called Springboard (4).

I did not collect self-identified race or gender, because I judged that asking about these identities in a considerate way could anchor participants or hint that representation was of interest in the study. In a future study, I would ask for permission to send an anonymous post-interview survey and include demographic data in that questionnaire. However, my sample appeared to be somewhat diverse: several participants mentioned their home countries, including the United States, India, Mexico, Australia, and England. Three participants appeared to be women and nine appeared to have an ethnicity other than white. There were no Black participants in this study (and therefore no Black women, the demographic group most affected by performance gaps in facial recognition). If I had collected specific, self-identified data about race, ethnicity, country of origin, and gender, I could see whether those features correlated with ethical sensitivity for an issue related directly to race and gender; future work can address this question.

All participants consented to have the audio and screen-sharing recorded; one recording failed (P21DS).

3.3.2 Think Aloud

I approached my research questions with a think aloud experiment. The think aloud protocol is a method in which participants speak their thoughts aloud as they complete a task and offers insight into what participants attend to, as well as the opportunity to observe their process [60]. According to Ericsson and Simon, concurrent verbalizations are believed to offer stable and accurate reports of ongoing cognitive processes, but for the purposes of this study, even if we only got insight into how participants interpret and talk about their work, it is still interesting: speech about work is the currency of collaboration, training, and management.

Concurrent verbalizations are classified as Level 1, 2, or 3, with the amount of internal processing increasing with higher levels [60]. When participants do more processing, their verbalizations contain more interpretation and therefore more information. I used instructions from Ericsson & Simon [60] designed to prompt level 1 and 2 verbalizations. These types of utterances offer information about what the participant is attending to and some more information from processing, but aren't thought to impede creativity, change decision-making, or alter the structure of task performance. There is evidence to suggest that level 2 verbalizations slow down task performance, so recorded times (i.e. time spent looking at Datasheets) will be compared between participants, but not assumed to generalize to real work environments.

Participants worked on their own computer, with their own software and settings. Previous studies of ES have relied on surveys and interviews, usually in reaction to written, hypothetical scenarios. This think-aloud method moves ethical sensitivity methods forward

by preserving some situational factors while still permitting researchers to control of key features of the ethical situation.

3.3.3 Materials

Problem Statement

I provided all participants with a problem statement describing a national chain jewelry store that wanted to first build a face detection model to collect data from their stores, and then a face recognition model to identify repeat offenders and suspicious behavior. The problem statement is provided here as Appendix E. This problem was selected because it has a variety of ethical issues that could be noticed and further investigated by participants, just as a real work situation could. Known ethical issues planted in the problem statement were: privacy for training data subjects, privacy for those at the jewelry stores, bias in facial recognition, and "suspicious behavior" detection as punishing pre-crime. As expected, participants noticed other potential ethical issues and offered nuance to the known issues.

Think aloud sessions took place in July 2020. The news and social media were discussing ongoing protests in the wake of the killing of George Floyd. Although the intent of this project was to write a problem statement with several potential ethical issues in order to get plenty of data about recognition and particularization, issues related to race and policing may have been more top-of-mind during the study period than they would otherwise be, especially for non-Black participants.

Data

I described the data to participants as "a random sample" from a larger dataset they were asked to consider to address the problem statement (To read the script, see Appendix B). I selected 171 images from the Flickr-Faces-HQ (FFHQ) dataset [107]. FFHQ includes

faces "in the wild" and is well documented. To manufacture demographic imbalance that could cause a biased performance gap, I intentionally oversampled images of people who appeared to be male and who were light-skinned. A sample of the provided data as participants saw it in Google Drive is provided as Appendix F.

The data set presented to participant appeared [1] to be composed of 71% images of men, 24% images of women and 5% images that were either not clear or contained people of more than one gender. 89% of the images appeared to be of white people, 5% who were not white, and 5% images of people whose face was obstructed enough that I could not classify their race. When selecting images for the unrepresentative data, I did not attempt to identify a person's specific race or ethnicity, but instead to ensure that people who appear to be of Western Eurasian descent were significantly over-represented. It may be that some of the images of people that I classified as white would not be classified as white by a given participant and that if each participant were to classify the data available in the same categories I did, that our classifications would not precisely overlap. Despite a few such hypothetical disagreements, I believe that the overall impression of the data as substantially more white and male would remain.

In addition, 27.5% of images included a person wearing glasses, 5.3% contained a person wearing sunglasses, 3.5% contained faces that were significantly obstructed, 7% of images contained more than one complete or partial face, and 15.8% contained a person wearing something on their head (for example, a hat, helmet, headband, glasses, over the ear headphones or headset). The dataset included one subject who appeared to have Down Syndrome and two subjects with dramatic costume make-up. Various ages were represented, including young children.

[1] These labels do not capture the self-understood identities of those in the images, nor the full range of race or gender groups, but rather a need to describe the extent to which the dataset was dominated by images of people who would appear to participants to be white and appear to be men.

Datasheet

The Datasheet intervention was designed by Gebru et al. [74][2]. I filled out the Datasheet with information from the original dataset's curators [3] and added fictionalized details for the purpose of the study. Fictionalized details included replacing Flickr with Photobucket as the source. Photobucket is a similar site for which user demographics are readily available to any participant who searched for them. Other details, like the exact number of images, were altered slightly so that the FFHQ dataset would not come up in an internet search of the provided details. I wanted to ensure that the original dataset wasn't associated with the experiment because it could muddy participant interpretations of data provenance and so the participants would not confuse this intentionally under-representative dataset with the original.

That Datasheet acknowledged two potential ethical issues explicitly. First, the data reflects the demographics of its source, which is heavily male and white. The Datasheet acknowledged this fact in the ethical considerations section. Second, the training data was said to be scraped from a website where users posted them with permissive Creative Commons licenses; the Datasheet admits that although the posters of the images were certainly aware that the images were public and had made them available for some uses, the subjects of the photos had not necessarily consented, but the document doesn't label it as an ethical issue. The other ethical issues planted in the problem statement (see: section 3.3.3) were not acknowledged in the Datasheet. For the arrangement and wording of the Datasheet, see Appendix C.

[2]The version used was included in the March, 19 2020 update of the paper available at: https://arxiv.org/abs/1803.09010

[3]Provided in the readme.md file on github https://github.com/NVlabs/ffhq-dataset

3.3.4 Study Design

This project deployed the think aloud protocol in two groups: 11 of 23 participants received the dataset, problem statement, and Datasheet, while the other 12 were issued only the problem statement and data. Participants were randomly assigned between the two groups.

I asked participants to explore the materials and formulate a plan to address the problem. Participants were able to view the data in Google Drive or download and work with it in software of their choice and were asked to verbalize their what they attended to as they worked (for specific instructions and prompts, see Appendix B). They were asked to consider "whether and how" to apply the data to the stated problem for 25 minutes, and to have a plan at the end of that period for what they would do next. Their screen and audio were recorded, and the verbalizations were transcribed by the author. Avoiding interrupting participants as they spoke, participants were asked to stop working after about 25 minutes. Several participants naturally concluded earlier, offering a summary of their next steps before being asked, and a few reflected and searched for longer. The duration of participants' think-aloud sessions are available in Figures 3.3 and 3.2.

After the think-aloud session, I asked questions using a funnel-sequence interview [116]. Inspired by [172], I used the funnel sequence to classify recognition into three time categories. Categories allowed me to capture as much recognition as possible before revealing the topic of interest and to capture recognition that perhaps happened, but which participants thought wasn't relevant to the study. The interview started with questions summarizing and clarifying the participants' plan: "Can you describe your approach?" and "What would your next steps be?" "How would you approach labeling?" Then, I wanted to to elicit any limitations of their plan participants were aware of: "What would an ML model trained on this model be useful for or not useful for?" Question 6 is even more direct:

it asks about one possible mitigation for some ethical issues in a flawed dataset ("Would you want any other kinds of data to improve the model?") Finally, Question 7 asks directly: "Did you notice any potential ethical or legal issues in the problem or data?"

Swenson-Lepper described the three time categories used in funnel sequence interviews clearly [172]: Time A) participant relates their perception; Time B) participants are asked about moral aspects of the situation without being directly asked about ethics and; Time C) participants are asked directly about ethics.

I recorded whether each participant's first recognition occurred while they were thinking aloud (analogous to Time A), during the interview before the direct ethics question (Questions 1-6, analogous to Time B), and during the interview as a response to the ethical question (Question 7, analogous to Time C).

3.4 Analysis

After the think aloud sessions, I used an automated transcription service (otter.ai), then listened to the audio while reading the transcripts to correct them. Then, I read the transcripts while watching screen recordings of the think aloud sessions. This allowed me to note moments of recognition, particularization, and judgment as well as what participants were seeing on their screen at the time. I recorded the time each document was opened and the time participants navigated away from it during the think aloud session. I used working definitions of each type of verbalization to label them (in Nvivo 12) and collect the time those verbalizations began.

The Datasheet was of particular interest. I labeled any comments participants made about the Datasheet, which parts of the Datasheet elicited reactions, notes, highlights, or other reactions indicating significance to the participant.

I plotted unprompted recognition, particularization, and judgment verbalizations, along

with screen contents throughout the think aloud session, in Figures 3.2 (with Datasheet) and 3.3 (with no Datasheet). I also labeled any recognition, particularization and judgment that occurred after the think aloud session (during the interview) as being "prompted" ethical sensitivity verbalizations (the interviews and these prompted verbalizations are not reflected in the timelines).

Recognition

The first time a participant mentioned a particular ethical issue in relation to the task they were working on, I recorded the time the utterance started, relevant comment text, and screen contents.

Cues were often clear, but not always. Frequently, participants read snippets of their screen contents aloud, highlighted text, pointed to things with their mouse, or paraphrased study materials as part of the recognition utterance. When participants did none of those things, I could tell what they had on their screen, but not what they were looking at. I had planned to compare my assessment of cues with participants' responses to an interview question, "What caused you to notice [ethical issue]?" Participants struggled to respond to this question, and got general answers about how participants had heard about the ethical issue or what they knew. I tried some re-worded versions of this question (e.g., "What tipped you off?") to no avail.

Whether an event was recognition or not was not always clear. Several participants would, for example, mention an example of facial recognition being used for an ethically-problematic task without explicitly voicing an objection (I did not count this as recognition) or describe facial recognition as "scary" (I did count this as recognition). Fortunately, all the participants with these ambiguous utterances later exhibited clear, unprompted recognition. This means that these instances didn't change the overall count of unprompted recognition,

but did mean I wasn't confident enough to report time-to-first-recognition as a meaningful measure. This ambiguity could be a weakness in this way of measuring recognition or it could indicate that recognition is not in fact a single awakening moment, but can also be a dawning realization.

Particularization

For the purpose of extracting particularization-related behavior and utterances, I developed the following definition of particularization behavior, based primarily on Weaver et al. [184, 186] Blum [24] and ongoing work:

> "Seeking information, reflecting, and making developmental evaluations about the situation, stakeholders, consequences, responsibility, options, resources, and the relationship of the issue to the task. Building understanding, connecting personal and external ethics to the details of the situation."

While coding transcripts, I labeled these utterances as "particularization," and recorded the time, notes about context, and screen contents. I also classified particularization utterances by the names of issues listed while coding for recognition. If the participant was seeking information or citing information from memory during particularization, I recorded the source and topic of external information.

Judgment

Although the post-task interview asks directly for a judgment, some participants offered judgments as they worked. Rest defines judgment as: "formulating the morally ideal course of action; deciding what one actually intends to do; or executing and implementing what one intends to do" [148].

I labeled any statement as a judgment which a participant considered a course of action. I anticipated that this would be difficult to disambiguate from instances of particularization in which people mentioned options, but participants nearly always used phrases like "I would" and "we should" (or a hedged version, like "I probably would") during the talk aloud session. I recorded both prompted and unprompted judgments, along with the time (for unprompted judgments). Judgments were classified by issue name and by the plan of action.

3.5 Results

This chapter sought to answer two research questions:

> **RQ 1**: Are engineers who read Datasheets cued to recognize ethical problems differently or at different rates than those who do not read them?
>
> **RQ 2**: What information on and off the Datasheet do ML engineers use to particularize a perceived ethical problem?

In reporting these results, participants will be referred to by a participant number followed by a letter indicating whether they were in the group that got a Datasheet ("DS") or the group that did not ("N").

3.5.1 Recognition

Figure 3.4 shows the first mention of an ethical issue by participants with and without Datasheets and whether it happened unprompted (during the think aloud session), in the interview before the direct ethics question, or in the interview in response to the ethics question. One participant did not mention an ethical issue at any time. Table 3.4 show what each participant had on their screen when they noticed the ethical issue.

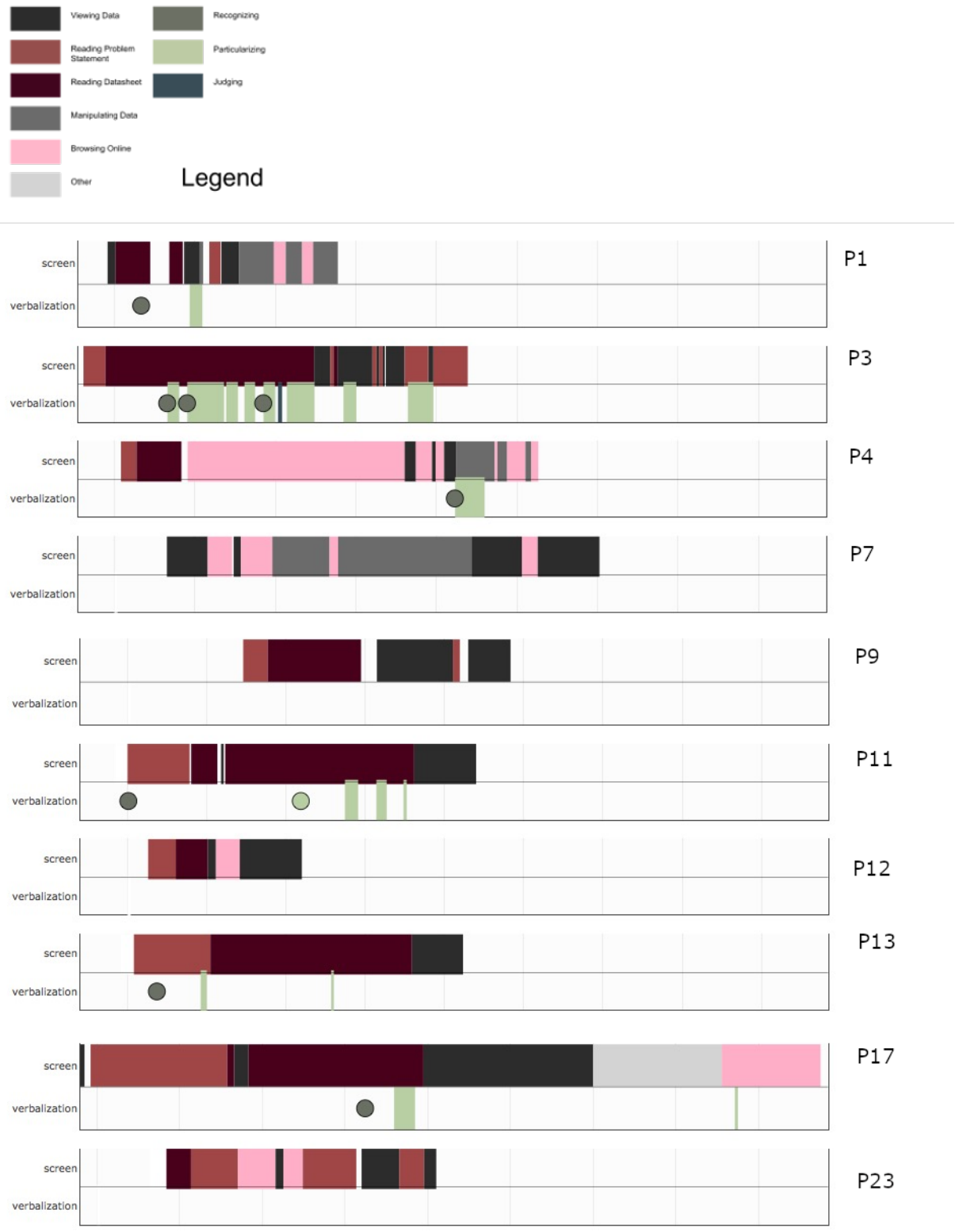


Figure 3.2: Screen contents and ES verbalization for participants with Datasheets

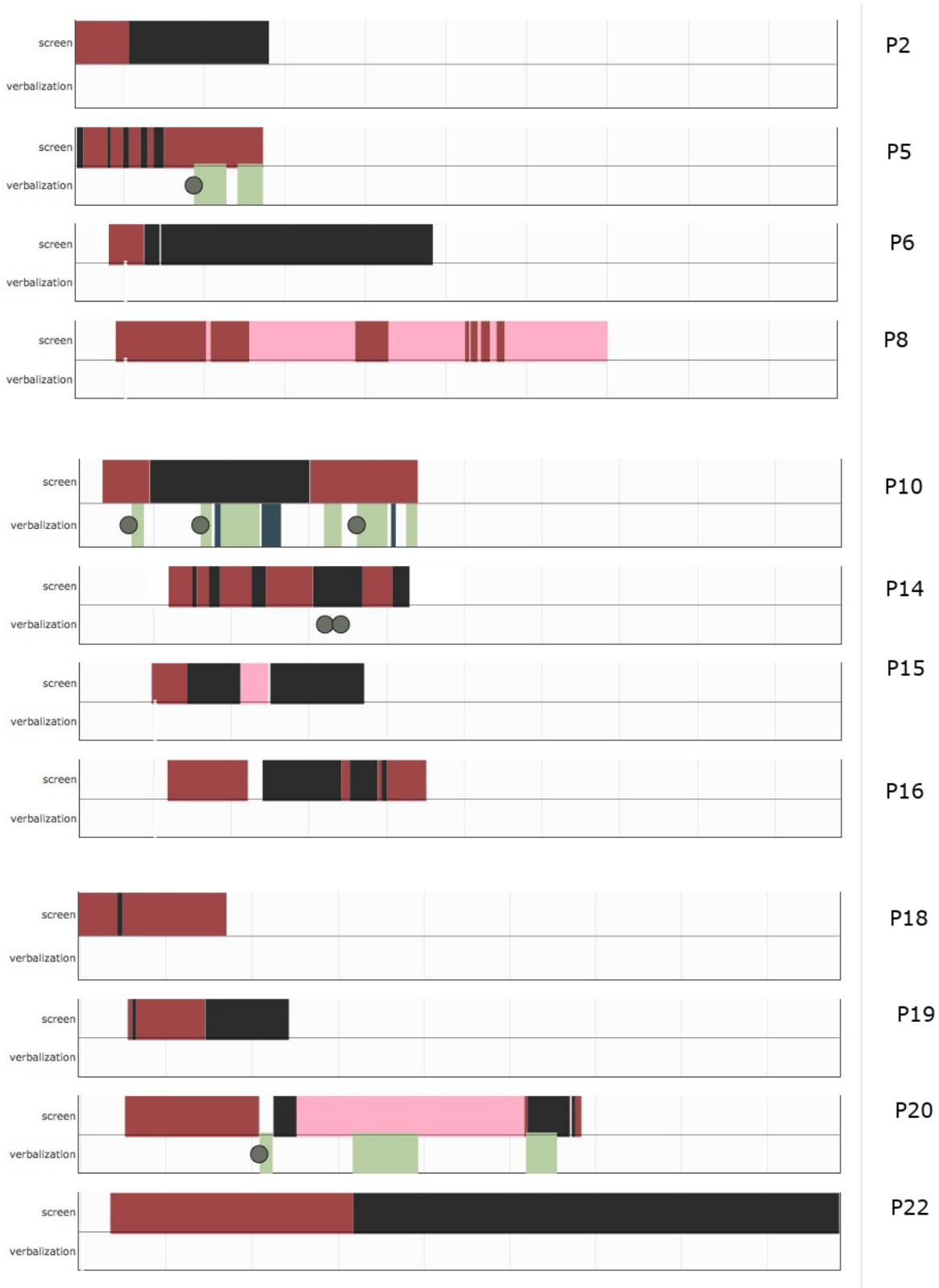


Figure 3.3: Screen contents and ES verbalization for participants without Datasheets

Of the 12 participants who did not receive a Datasheet, four mentioned their first issue unprompted and eight first mentioned an issue in the interview: three before the ethics question and five after. Three of the unprompted participants were looking at the problem statement when they mentioned their first issue and one was looking at the data.

Out of 11 participants who received a Datasheet, 10 read at least some of the document. Of those, seven recognized at least one ethical issue during the think aloud session and three mentioned the ethical issue during the debriefing interview, one before the ethics question and two after. One participant in the Datasheets group did not read the Datasheet did not mention any ethical issues, including in response to direct question about ethical or legal issues in the task.

Four participants (P1DS, P3DS, P17DS, and P21DS) mentioned their first ethical issue while reading the Datasheet.

P3DS and P21DS brought up a privacy concern while reading an answer about confidentiality which acknowledged that the subjects of the photos may not have consented to their publication:

> "So the subjects didn't give permission?" *-P21DS (User highlighted confidentiality question and answer. The recording for P21DS failed: the source of this quote and context is handwritten notes.)*
>
> "Yeah, I'm a little concerned about, you know, some of the ethical implications here as well of downloading a whole bunch of people's images and then using them for this study that might not have expected." *-P3DS (User tracked mouse over text as they read)*

P1DS and P17DS mentioned bias while reading about the confidentiality and selection of the data:

Table 3.4: Cues for first-issue recognition

Datasheet		No Datasheet	
Cue	**Participants**	**Cue**	**Participants**
Datasheet	4		
Data	1	Data	1
Problem Statement	2	Problem Statement	3
Interview (Prompted)	3	Interview (Prompted)	8
None	1		

> "So random selection of those, good. Wonder what different what the variety or span of that was. Okay, so it came from [pause] where's it come from? Photobucket. Okay. Okay, so whatever the demographics are for Photobucket, that's what I can expect in here. And I imagine Photobucket allows non facial images, so it'll be interesting to know how they decided whether it was a face or not, in order to create the labels that they had? Perhaps they did it with humans, perhaps, or used a pre-trained model and that could introduce errors in the dataset— biases." *-PIDS (User tracked mouse over text as they read)*

> "Data was sampled randomly. Hm. I wonder how they did this demographic representativeness bit . . . we're dealing with image processing, which often has trouble with skin tones. So kind of leads to racist machine learning more or less." *-P17DS (User read some words on the Datasheet aloud and highlighted the phrase "basic demographic representativeness")*

Ethical issues mentioned by participants included: discrimination from demographically unrepresentative training data (15), high stakes in facial recognition (particularly for false positives) (7), privacy and consent in provided training data (9), privacy and consent in data collected from the store (5), other privacy concerns (2), unconscious bias in law enforcement or security personnel (1), and justice implications of predicting crime and acting on those predictions (i.e. "Pre-Crime") (1).

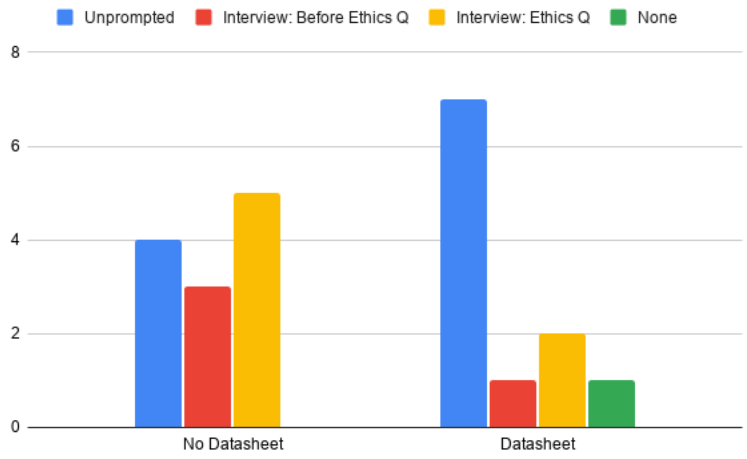


Figure 3.4: Prompted and unprompted issue recognition

2*Issue name	Without Datasheet			With Datasheet		
	Think Aloud	Interview	After Ethics Q	Think Aloud	Interview	After Ethics Q
Biased Training data	4	1	2	5	2	3
Privacy in Training data	2	2	2	2		2
Privacy in Use	1	2	2	1		
Other Privacy Concerns			1		1	1
High Stakes	1	2	1			2
Unconscious bias	1					
Pre-Crime	2				1	

3.5.2 Particularization

Participants who particularized out loud while they worked spent a widely variable amount of time doing so, ranging from 33 seconds to 9 minutes and 17 seconds (average of 3 minutes and 52 seconds). One participant who recognized the issue unprompted did not particularize out loud at all.

I anticipated participants would seek out information about the ethical issue primarily; in fact, they built and reflected on broader technical and social understanding. Particularization utterances revealed that participants' understandings differed substantially.

Participants built and relied on their technical and social understandings while particularizing. While developing and exploring both types of understanding, participants reflected on their existing knowledge and beliefs, sought (or indicated that they planned to seek) information, and relied on the study materials. Collectively, they used eight examples of past engineering failures and one of an engineering success and made trade-offs (between gains and risks, benefits and costs). Using both ethical and technical understanding, they discussed their options for mitigating the ethical issue, considering resources and risks.

Social Understanding

While deciding what to do, participants considered the perspectives of and relied on their beliefs about a variety of stakeholders, including data subjects, shoppers, thieves, the business implementing the system, and law enforcement. The example of law enforcement will show how differing views about a group of stakeholders shapes participants' view of the morality of the system and its potential ramifications.

P11DS used their beliefs about and relationship to law enforcement to support their moral evaluation of this project as part of particularizing.

> "In general, as a law abiding citizen, I am interested in supporting law enforcement . . . So this [project targeting thieves] is this is acceptable at a moral level. Now, if you wanted me to do a face detection to, you know, detect something racial with regards to admission to universities than I say, uh, nuh-uh." - *P11DS*

P3DS had a view of law enforcement that led them away from moral approval. P3DS didn't consider their own relationship to law enforcement, but used their beliefs about police behavior in a hypothetical scenario:

> "I don't know, we don't want to get too much into politics here, I guess. But, you know, for certain, especially ethnic groups that might be able to police come into the store and they think, random person who they already suspect as a criminal, and so they're prejudiced against him. And now he's reaching into his jacket to pull up his wallet because he wants to buy a necklace, you know, for his wife, and . . . the police see him reaching in and pulling out some like black object and think it's a gun like they're, you know, could potentially be like, serious ramifications" *-P3DS*

P5N had initially approved of using the system only to catch repeat offenders. They used beliefs about law enforcement's use of data to reconsider.

> "So now I do think that the third task with the repeat offenders, after [mentally] processing a little more, would also be a little bit concerning. Like, how do you even make that data set repeat offenders right? It's still probably like a police institution. So that would also be biased and you're more likely to catch sort of stereotyped individual more than others." *-P5N*

Technical Understanding

Participants relied on and sought information about the data, the technology, and the deployment circumstances. Among these, there were areas of consensus and conflict.

Participants differed in their views of some essential training data composition questions including whether, in order to be useful for the problem statement, the data provided should have or not have negative examples– images of things other than faces– or images with more than one face in them.

P16N and P20N talk about removing images with more than one face, framing it as "data cleaning:"

> "Not this one. Because this one also has two faces. I think we have we have to make sure the data is clean before we put into the model." *-P16N*

"I have one whole desktop designated to data cleaning. So we're looking for how do we remove blurry pictures? How do we remove you know, pictures that are have multiple people in them or or maybe only half a person . . . let's remove those so they're not kind of adding any noise." *-P20N*

Most other participants agreed that images with multiple faces are necessary. P17DS explains, using the Datasheet to support his understanding of the data:

> "In real production, security cameras, you're gonna have more than one face in a bunch of images. All the images [trails off, reading Datasheet] oh, 'is centered on the images center pixel.' Okay, so that's as I feared. So we're gonna have to deal with for an actual production situation, we have to deal with not just detecting a face, but also like, sort of centering things if we're going to do a more like machine learning approach here." *-P17DS*

In contrast to the disagreement about negative examples, there was widespread agreement the difference between the provided dataset and the actual deployment scenario was going to be a problem to contend with. The following quotes illustrate how this technical understanding operated within particularization to generate ideas about what to do next.

> "Geez, how do we deal with the problems between these two datasets? Because they're going to be different, right? You know, like, we're gonna have off the shelf security cameras versus whatever– these look like decent photos. Maybe we can like white balance the photos and then do black and white and have them like sort of cropped to the face so that they're like kind of close." *-P17DS*

> "So, the more I think about it, the more I think I probably would want to skip this step to be honest. I mean, I want the data from the store." *-P13DS*

Several participants framed the problem as multi-stage and identified some stages as possible with the current dataset, analyzing the feasibility in turn. P18N offers an example of this:

> "But then in the last statement when they also say that they want to use this model to predict suspicious behavior, so that's, that's another step. So for that, like I said, one would need labeled data . . . And this, this suspicious behavior just cannot be inferred from their face, I think this would require more like tracking of the path that customer– where you went in the store, who you talked to, how much time is spent where, stuff like this. That is much more complex problem, because it will require a lot of generation of training data for this person within the video frame such as like a time series feature. Not sure how feasible it is to do that. It will require lot of training data." *-P18N*

Problem, Options, Resources, and Risks

To understand the problem, its sources, effects, and what can be done about it, participants synthesized social and technical understandings. In other words, the ethical issues in this study are sociotechnical problems [155].

P3DS relied on their understanding of relevant technologies with beliefs about user's perspectives to formulate an understanding of the ethical aspects of technology design choices.

> "So if I like make a new commit where I got rid of, you know, John's awkward photo, his bachelor party, then John's photo is still there, if anybody has a link to the previous commit. So I would raise that issue with whoever is the guy responsible for maintaining this. And the GitHub is probably not where you want to store this, if you want to be able to, like, have a revised data set. I mean, in general, once the data is out there, you know, anybody who's downloaded it already has John's embarrassing photos still, but GitHub, it's even worse because it's even there on GitHub." *-P3DS*

> "Sure, it was all public. People still don't necessary assume that everybody is going to be able to access this stuff even when they make it public. Some people don't realize they know the default settings often, like everybody in the world can see this and you go somewhere to change that. So they, you know, might have not realized they were opting into us collecting their information." *-P3DS*

Like social understandings and technical ones, significant differences among participants' perspectives emerged when addressing sociotechnical issues as well, as seen in these three participants' thoughts about the risks of false positives.

P20N considers the costs of false positives ethically and financially.

> "You know, this isn't this isn't holding a kid while they're, you call their mom ... This is people's lives are potentially on the line. And is that worth saving 500 dollars for a stolen ring? ... Or even you know, if you want to be callous and think from a business perspective, is the backlash for an individual having the cops called and an incident happening and the loss of revenue from people boycotting your score because from your store, even if you don't care about About the ethical side of it." - *P20N*

> "So, what happens if your model says something and it's wrong? I think it's the main thing ... I want to know. How bad can that be for a person? And if it can be really bad, then you have to seriously consider, like, is using a model here going to provide benefit or not? And how can we like make sure there's sufficient human in the loop involvement?" - *P3DS*

P23DS compares the risks of false positives with the costs of false negatives and arrives at a very different understanding.

> "The main goal will be true positives, and at least in these types of situations, I guess it's the goal is to catch as much as possible. So, not a problem if it's false. It's, it's, I think it will be, we will tilt a bit more to okay if we have more false positives than if we have false negatives, in this case, some thief appearing and not being identified will be more costly." - *P23DS*

Datasheets and Particularization

In the group with Datasheets, most particularization happened while the Datasheet was on the screen (see Figure 3.2). In some cases, reading the Datasheet guided participants through particularization. Two very different examples of this are P3DS and P11DS.

Participant 3

P3DS did the most particularizing of any participant (9 minutes and 17 seconds). They first recognized ethical issues while reading the Datasheet. The Datasheet's existence reassured P3DS somewhat ("Maybe that means some of the concerns about the data use have obviously been aired"), but they still read it in detail and engaged with it critically.

After reading a section about Creative Commons licenses, P3DS felt somewhat assured, but they didn't resolve the concern entirely:

> "Or at least legally, we look like we're okay even if whether we're okay, morally slash ethically might not be the same question."

They continued to look at the Datasheet with a critical eye:

> "It might be nice if they said why it was deleted to see, you know, anytime there's like a bias you're inserting in your data, right? You want to know like, what was that bias? ...I'm assuming maybe people just flagged those as offensive since it mentioned you could do that."

Throughout reading the Datasheet, P3DS engaged with the socio-technical problem at hand in detail, including considering the implications of using GitHub for user privacy (See section 3.5.2).

While reading that "each instance includes at least one human face" in the Datasheet, P3DS again used technical understanding to reason about social issues:

> "Probably ask whoever sent me this how they determined that like did they have a pre-trained model that is already good at face detection? Or is this like a person went through manually and said like face no face for each one? Just to wonder like, are you getting some sort of bias here in terms of you know, you only have images that have easily recognizable faces because something already recognized that there was a face in."

Participant 11

On the other hand, P11DS initially focused very directly on the technical aspects of the task. Although they nearly instantly acknowledged the possibility of an ethical issue ("A face detection algorithm. Those are fun and scary and all kinds of nice things" at 51 seconds into the think-aloud session) and read the entire Datasheet in order, they jokingly dismissed much of the content as not part of the task at hand.

> "Archive, whatever. Restrictions, something, I don't think I care. Okay. For this purpose anyway. Confidential, Karen [the author] will take care of it, don't care."
>
> [*Reading*] "Identify these subpopulations [*trails off*] to possibly identify individuals. Yeah, sure. But do I care? Do I care? No, no, really don't care. Face recognition: it's a generic face don't really care who it is just that it's an actual person."

However, after reading some more, the participant reported a paradigm shift, not unlike what's described in the ethical sensitivity literature as an awakening [184]. Here P11DS paraphrases one of the questions, reads the answer, and responds.

> "Has an analysis of the impact of its use on the subjects been conducted? No. Alright, now I'm starting to feel uncomfortable. Maybe? [laughs] Yeah. So maybe you– maybe my using these these faces is . . . it's public, but if the security cameras . . . I don't know, it's something private. And I'm starting to think other thoughts here beyond the immediate task at hand."

Although P3DS and P11DS approached the task and Datasheet with a very different sense of the relationship between the ethical issues and the task at hand, both demonstrated high ethical sensitivity with unprompted recognition and particularization, and both used the Datasheet to shape their perception of the ethical aspects of the data.

Particularizing without a Datasheet

I have much less data about how participants who did not see a Datasheet particularized. Only four participants without Datasheets recognized during the think-aloud, three of whom particularized.

P5N particularized for 3 minutes and 37 seconds (longer than average) and spent that time reflecting, citing an example of an engineering failure. They described where they understand bias in ML to come from:

> "Well, you can try to train a network and do anything you want. But there has to be sort of pattern. And I can, I would argue that there's not necessarily a pattern between someone's face in suspicious behavior in stores. And of course, there's, like in this current political climate, I mean, there's really bias and everything. So there will definitely be bias in your training data for this fourth stage. For example, certain types of people will be represented more often in the training data, just because of implicit bias."

P10N particularized for 7 minutes and 51 seconds (much longer than average), most of which was spent reflecting on the circumstances of use: the behavior of thieves and innocent shoppers, the setting of stores in malls and how security works there, and the relatively diverse demographics of the U.s. Notably, P10N mentioned the "Coded Bias" project almost immediately– before they saw the data. P10N then opened the data and noted that they thought the data was "almost uniformly distributed."

P20N particularized for 7 minutes and 41 seconds (much longer than average). They discussed the circumstances of use as well, including the behavior of innocent shoppers and thieves, and comparing the context of this project to the context of projects they have worked on. P20N also discussed the incentives of the store:

> " And also, you know, it's a company public image. If it comes out that a jewelry store is removing all males between 20 and 24 who are in . . . a certain minority group and then that really is going to impact sales a lot more than that . . . Like some companies would rather just have the thefts that you can write off than actual loss of revenue from– from being racist, sexist agents, etc. . . . [that's something] particularly with machine learning, you can get a lot of backlash for. So that's something I'm always looking for, both from an ethical perspective, but also it's a business when I'll make money, we'll make sure we're not losing the money from it."

3.5.3 Judgment

As I expected, participants did not do as much judging during the short think-aloud session as they did particularizing.

The small amount of unprompted judgment makes it difficult to compare judgment with and without a Datasheet, but this study did offer data about ML engineers' judgments about the ethical issues in facial recognition training data.

The interview prompted judgment among seven additional participants (five with datasheets, two without). Participants discussed ways to move forward to mitigate privacy threats, high stakes from false positive predictions, and biased training data. Several participants considered actions that would mitigate more than one ethical issue: P3DS, P13DS, P17DS, and P11DS considered seeking out a different dataset to replace the one provided. P10N considered stopping the project altogether. P3DS suggested supplementing or replacing the training data with data created by employees, citing the example of TSA employees creating training data for a Kaggle competition. P11DS and P4DS mentioned that speaking with the company's legal counsel or department would give them some peace of mind.

Although many participants expressed concern about privacy and consent, only two participants considered judgments to mitigate these concerns. Several participants suggested broad interventions that would address more than one issue: putting a human in the loop (P20N, P14N, P13DS, P10N, P6N, and P3DS) and seeking out a different dataset to replace the one provided (P3DS, P13DS, P17DS, and P11DS). Most participants considered actions to mitigate the bias issue in the training data.

Because I asked about it in the interview, most participants considered actions to mitigate the bias issue in the training data. The most popular solution mentioned for dealing with bias was altering the demographic distribution of the training data. However, these ideas differed on two dimensions: the goal and the means (See 3.5 and 3.6. Participants often changed their minds here and their ideas did not align others'.

P6N and P11DS both considered making training images darker. P6N pursued this angle, searching for methods to darken skin tones without darkening the background as well, and planned to automate the process. P11DS changed their mind as they considered this option, an example of a common pattern in judging:

> "To me there are two approaches. One approach is to get sufficient data in the lacking areas to fill it out so that there is a better representation. So hey, go back to [the client] and say hey, is there is there any additional etc., or somehow do some photo magic and [*pauses*] create [*pauses*] skin tones on– yeah, I don't know, create skin tones, but then facial features are different too. So that's probably not a great idea." *-P11DS*

P9DS brought up a similar technical solution, but didn't consider it.

> "Let's say the image is in really good lighting. And you want to give your algorithm some examples where the lighting isn't as good. Right? So that's something that you can basically, you know, apply a filter . . . And that seems

Table 3.5: Goals of altering demographic distribution

Goals	Participants
Demographic distribution reflects jewelry store location or customer base	P4DS, P13DS, P14N, P17DS
Data is equally distributed across each group	P8N, P10N, P15N, P16N
Demographic distribution matches US distribution	P8N, P17DS
Demographic distribution matches criminal population	P5N
Demographic distribution should be such that accuracy is similar among groups	P3DS, P13DS

> reasonable to me, but I would not be comfortable. If we're like missing skin tones, or if we've got an imbalance between men and women. Like you know, it doesn't have to be perfectly representative of the population. But I don't know of a way you can do augmentation techniques to fix that you need actual – pictures . . . I can't just somehow take my pictures of men and you know, I mean there's there's research stuff, right. Like they mentioned that they even use this train a GAN. So theoretically, right. I could create a GAN and I could make more women. But, but I wouldn't be comfortable with that, I think it would, it would leave a lot of a lot of questions, a lot of unknowns."

The diversity of judgment on this issue reveals differences in understanding that those judgments are built on. The pattern of participants revising their judgments after further reflection demonstrates the non-linear relationship between particularization and judgment [147].

3.6 Discussion

Participants in both demonstrated ethical sensitivity. Although participants in the group who were given Datasheets had more unprompted recognition, suggesting that the Datasheet aided with recognition, all the participants in the group with no Datasheets recognized at

Table 3.6: Means of altering demographic distribution

Means	Participants
Collect more data	P2N, P3DS, P4DS, P7DS, P9DS, P13DS, P19N, P20N
Reweighing	P4DS, P13DS, P15N, P18N
Oversample from minority groups	P3DS, P11DS, P16N, P18N, P19N
Undersample from majority groups	P16N, P18N
Data augmentation (not specific)	P9DS, P14N, P18N
Use a GAN to generate more data from minority groups	P10N, P9DS
Darken existing images of light-skinned people	P6N, P11DS
Delete images for which algorithm doesn't work well	P7DS
Use a toolkit (e.g., IBM, Microsoft)	P9DS
Do more testing	P18N, P9DS, P15N

least one ethical issue after being prompted by interview questions. While prompted recognition on its own isn't helpful in a work setting, it gives us some guidance for developing interventions that serve as prompts (see Section 3.6.2.

P5N recognized the most issues out of all participants: three unprompted and an additional one during the interview. P10N recognized the bias issue within 90 seconds of starting the think aloud session, particularized extensively (for 7 minutes and 15 seconds), and referred to and accurately summarized "Coded Bias," a project by Joy Buolamwini about bias in facial recognition [32]. Ten out of eleven participants who were given a Datasheet read it. Together, these facts suggest that machine learning engineers exhibit ethical sensitivity and that the introduction of cues for recognition could be fruitful.

The difficulty of locating recognition in a single moment of "awakening," which the literature led me to expect [186] and the difficulty participants had answering questions about cues suggests that, at least in this context, ML engineers may experience recognition as a more gradual revelation.

3.6.1 Context Documents and Ethical Sensitivity

This study suggests that context documents in general and Datasheets in particular support ethical sensitivity among machine learning engineers working with unfamiliar and ethically problematic datasets.

The headline findings are good news for the authors of Datasheets and other context documents who hope that their intervention will raise awareness of ethical issues. More participants in this study with Datasheets mentioned ethical issues while working with unfamiliar, ethically problematic data than those who did not. Participants relied on them extensively to particularize, with most particularization in the Datasheets group happening while looking at the Datasheet. Although it's tough to evaluate ethical judgment without declaring some judgments better than others, four participants suggested replacing the dataset entirely, all four of whom had a Datasheet. Perhaps having more detailed information about data context and provenance gave these participants the confidence to make a call about the suitability of this data.

It's possible that participants in both groups recognized at the same rate as one another, and that whether a participant mentioned an ethical issue during the think aloud session is not a good measure of whether they noticed one. Perhaps participants who didn't run across discussion of the context of training data didn't think that the context was relevant to the task at hand. If the Datasheet has the same effect in the workplace – signalling to ML engineers that data context and ethical aspects are relevant– it is still achieving its aims.

When I proposed this study, I was concerned that participants would not read the Datasheet, so I had a plan to re-balance the groups if needed to ensure I had enough data from participants who opened the Datasheet. To my surprise, 10 out of 11 participants who were offered a Datasheet opened it with no encouragement. Three participants opened the document, exclaimed that it was long, and navigated away to view something else, but all

three eventually returned to it when they had questions about the data. Given the knowledge that document length could be overwhelming, though, authors of context documents may consider making them more brief, offering an outline or linked navigation, or highlighting important sections that they want to ensure people read.

The Datasheet prompted six recognition events, four of which were the first ethical issue a participant mentioned. Half of these occurred when reading text about something technical (e.g., recognizing a bias issue while reading about data selection.) This suggests that Bender & Friedman and other context document paper authors may be correct in believing that surfacing information about data distribution and context can trigger recognition, even without including direct ethical questions or language. The fact that four participants (one with a Datasheet and three without) mentioned their first ethical issue in response to indirect questions (early in the funnel sequence) further supports the assertion that surfacing dataset characteristics and likely effects may prompt ethical engagement.

3.6.2 Other Cues and Tools

Besides context documents, organizations can consider developing other tools, practices, and policies or shaping norms to encourage recognition, support particularization, and guide judgment. This study can offer some guidance for developing and evaluating these interventions.

Once asked about it in the interview, all but one participant either quickly mentioned or took a moment to consider whether there were ethical issues involved and were able to cite at least one. This suggests that a tool or policy that involves simply asking technologists about potential ethical issues may go a long way improving ethical recognition during development. Maybe better than consistent questions as part of a regular meeting or form, that could eventually prompt a habitual "no" are intermittent prompts: perhaps something

analogous to the experience sampling method [112].

Several participants said they would reach out to the company's legal department or counsel, and several more expressed the desire for a third-party ethics watchdog, rating agency, or review board. A source of independent advice may give technologists peace of mind, information that will help them recognize and particularize future ethical issues, and encourage them to feel more comfortable engaging with ethics in their work. P4DS put it concisely: "It's important to be able to raise your voice without losing your job."

Far from making engineers worry for their jobs when raising ethical concerns, a particularly strong intervention may be to design job descriptions and evaluations to include ethical engagement. Making it clear that noticing ethical issues is part of their work responsibility and rewarding that engagement with positive reviews, raises, and promotions could go a long way to ensuring that engineers are looking for and are willing to report potential ethical issues.

Think Aloud

This study was the first to use a think-aloud method to observe ethical sensitivity. Think-aloud offers some difficulties and advantages, but overall renders a very rich view of ethical sensitivity compared to existing methods.

Think-aloud really shines when it comes to observing particularization. Until now studies of particularization have been very inconsistent and acontextual, like asking participants to list and rank factors that they considered when responding to a scenario. Think aloud, even in a simulated work context, offers a rich view of particularization. It allowed me to watch participants search for information, use examples, rely on existing understanding, and reflect. It allowed me to see what existing understanding mattered to building understanding and how those understandings differed among participants. I believe that think

aloud will give us a more grounded and more complete conceptualization of particularization as well as a view into how it works in different circumstances.

Think-aloud also offers an improvement over existing methods when it comes to observing judgment. Rather than a selection or single statement of a participant's decision, think-aloud lets us capture the full range of judgment. The verbs Rest uses in his initial conception of judgment are "formulating", "deciding", and "executing or implementing." We saw quite a lot of this detail in judgment: we saw people explore options, change their minds, make trade-offs, and "if [*condition*], then [*judgment*], but if ..." A think-aloud study with a different scope or an ethnographic method could better include the "execute and implement" phase of judgment. None of this insight is available in surveys or other methods that focus on the ethical decision. Looking further into these developmental judgment activities may help us intervene into this key moment of technology development.

3.7 Limitations

This study demonstrated that think-aloud studies can be used to study ethical sensitivity in machine learning. However, recognition and cues were more difficult to observe than expected. Difficulty identifying recognition was detailed in the discussion. This limited my ability to compare time time to first recognition among participants and the average between the participants with and without Datasheets.

I did not collect self-identified race or gender for this study. In retrospect, this information could have offered useful context. Especially in light of national news events related to race, this context would have allowed better reflection on the standpoints of participants.

Further work can be applied to how to observe recognition and cues precisely during technology work. I encourage future work to be as highly situated in real work contexts as possible to ensure that we get an accurate picture.

3.7.1 Future Work

This study suggests several profitable avenues for future work. Better understanding ethical sensitivity across technology development contexts will allow us to intervene in that work to encourage recognition, support thorough particularization, and guide judgment. Researchers can continue to use think-aloud studies to study ethical sensitivity in new contexts, to test a variety of different interventions, or with more subtle ethical violations, especially when particularization and judgment are of interest. In addition to other context documents, it'd be interesting see whether interventions like envisioning cards [72], adversary cards[191], and design workbooks [192] elicit or change the character of ethical sensitivity.

All three of participants without Datasheets particularized for longer than average. It may be that the guidance of a Datasheet supports more efficient particularization but with only three non-Datasheet particularizers, I don't believe this study offers enough data to support the claim. Further study on particularization with and without context documents could shed more light. Altered or new methods can be developed to focus on recognition, to observe ethical sensitivity in groups, or to describe ethical sensitivity in action in real work settings.

3.8 Conclusion

This study suggests that context documents may prompt recognition, support particularization, and guide judgment in technology work. It demonstrates a method that renders rich insight into ethical sensitivity and how interventions aid or hinder ethical sensitivity during technology development. Using this method, this study offers a first look into ethical sensitivity in technology development and reports the most detailed, contextual view of

particularization and judgment yet.

This study shows an example of one part of attending to ethics in ML: interventions that encourage ML builders to notice and build understanding of ethical problems as they work. I believe that to effectively address the potential harms of this widely applied and quickly developing technology, as many people along the pipeline need to be engaged in the project of mitigating ethical issues as possible. Yes, user boycotts. Yes, citizen engagement. Yes, refusal to build. Yes, ethical interventions in training data, training, and post-training.

We need to know what helps workers notice, engage, and come to a decision all along the process, for subtle issues as well as issues in the news. This study offers encouraging evidence for context documents and introduces one method for describing the impact of other interventions into machine learning practices. I hope this study encourages more work on ethical sensitivity in technology development in general, and ML training data curation in particular.

Chapter 4: Designing Up with Value-Sensitive Design: Building a Guide for ethical machine learning development

Abstract If "studying up" can offer insight into the workings and effects of power in social systems, perhaps "designing up" will give us a tool to intervene. This chapter offers a conception of "designing up," proposes using the structure of Value Sensitive Design (VSD) to accomplish it, and submits an example of a guide to ethical mitigation strategies for machine learning (ML) engineers. VSD allowed me to prioritize the values and interests of lower-power indirect stakeholders (citizens, marginalized groups, and data subjects) while designing a product to be used by high-power direct stakeholders (engineers, their educators and managers) in an attempt to mitigate harms from ML systems upstream. The designed artifact is a filterable field guide for ethical mitigation strategies. It aids machine learning engineers who have noticed an ethical issue in understanding and matching the particulars of their problem to a technical ethical mitigation; may broaden its users awareness of potential ethical issues, important features, and available mitigations; and may encourage ethical sensitivity in future ML projects. Feedback from ML engineers and technology ethics researchers rendered several usability improvements and ideas for future development.

4.1 Introduction

When I use a recommendation algorithm to help me select the next show or song I want to hear, the user and the subject of the system-supported decision are the same. In other words, I make the final decision about what to watch and the decision affects me. When a social network or search engine decides which posts, search results, or ads to display, the data subjects, users, and decision subjects are (in aggregate) the same, but the decision is made not by the user, but the algorithm. The user does not know what other options there were, or how ordering decisions were made. Lack of transparency takes some power away from the user.

Other ML-driven software is built for one party to make decisions that about others. Power differentials are common between the decision-maker and decision-subject in these cases: for example in medicine (including genomic, diagnostic, and mental health data) [39], law enforcement (including crime prediction [158] and parole evaluations [14]), employment (including hiring and evaluation [136] and credit [136]. As AI systems proliferate, existing power relationships over who can get on an airplane [173], who gets laid off or promoted [53, 136], and other decisions we allow one party to make about another will be reified and supported by technology.

Unfair, opaque, and invasive machine learning (ML) sometimes results from (e.g., [158]), and other times simply reifies (e.g., [39]) existing power dynamics. Harmful ML has inspired a lot of design to mitigate that harm. Governments, organizations, researchers, and advocates have designed policies, ML techniques, educational campaigns, and other technologies to be used by governments, organizations, researchers, users, and citizens to reduce harm, liability, and to protect these human values for their own sake (see reviews of guidelines [81], education [154], technology [55, 70]).

So how does the policy, tool, technology, curriculum or campaign you are designing

disrupt, support, or otherwise relate to relevant power structures? What is the relationship between you, as the designer, and the rest of the power dynamic?

The variety of ML ethics interventions is evidence of a promising willingness to intervene throughout the ML ecosystems: at different moments, using different means, targeting different people, and using different arguments. Empowering citizens and data subjects to try to protect themselves with education, browser plug-ins, and the ability to opt out is important, but leaving it up to individuals is not enough. A wide-net, "yes, and" approach to ML ethics will improve the chances that we catch and mitigate any given novel threat. This chapter proposes and demonstrates a method for designing for high power actors while mindfully mitigating risks for low power actors. It employs Value Sensitive Design [71] to the project of "designing up" [15] for machine learning engineers.

The design goal in this project is a digital field guide for ethical mitigation strategies in machine learning: a search tool that allows ML engineers to find ethical mitigations that fit their circumstances and goals; surfaces key aspects of fit to make building an understanding the nature and needs of ethical ML situations more efficient in the future; and introduces ML engineers, educators, students, managers, and researchers to the broad range of ethical ML research and design. It can be filtered by key aspects of the ethical problem and its technical context; each mitigation strategy has a profile that describes in more detail key features and links to content sought by engineers; and users can participate and expand the project by suggesting edits, submitting tool profiles, or viewing, downloading, and building on source code. It targets key harms created or propagated by ML– including privacy threats, outcome unfairness, procedural unfairness, lack of diversity, and lack of transparency– and intervenes with high-power actors who are upstream enough to mitigate harm.

4.2 Background

The aims and methods of this project were inspired by "designing up," structured by ethical sensitivity (especially particularization and judgment), and accomplished using value sensitive design (VSD).

4.2.1 Designing Up

Laura Nader encouraged anthropologists to not only study groups that are lower power in a social system, but also those in middle- and high-power positions. She notes that people with high power in social systems have broad public impact and responsibility [131]. Studying those in power in social systems allows us to "flip" our questions, Nader points out: "Instead of asking why some people are poor, we would ask why other people are so affluent?" This allows us to understand and critique power in social systems.

Barabas et al. argued for a similar reorientation in data science [15] . In their case study, they executed a similar flip: rather than studying re-offense risk of prospective parolees (a project noted for its racial bias [14]), they focus on judges and judicial culture. There is a substantial power gap between those who design and build algorithmic systems and the data subjects, users, and the sometimes-unaware citizens about whom models make decisions. They argue that data scientists who study up "could lay the foundation for more robust forms of accountability and deeper understandings of the structural factors that produce undesirable social outcomes via algorithmic systems."

This project hears the call for designing up *using* data science, and reflects it, designing *for* data science development. Inspired by Lilly Irani's encouragement to use design to intervene "upstream" from harm [97, 98], I will use Value Sensitive Design (VSD)[71] to design a tool for data scientists and ML engineers to make it easier for them to employ ethical mitigations. VSD includes empirical investigations, which allow us to *study*

up– with the aim of better understanding the workings of power in the design of influential technologies– and technical investigations that will let us *design* up– with the goal of intervening in the early stages of ML development.

4.2.2 Ethical Sensitivity

To support the values of lower-power actors in social systems that include potentially harmful ML algorithms by mitigating them upstream in ML algorithm development, this project aims to help ML engineers understand the particulars of and make decisions about the potential ethical problems in their work.

To conceptualize these two goals, I use the ethical sensitivity (ES) framework. ES describes a worker, focused on the technical aspects of their task, who experiences a paradigm shift (*recognition*) when they realize that the task at hand may have ethical implications. They then reflect and seek information about their situation: the details of the circumstances, the opinions of relevant actors, stakeholder interests, relevant internal or external standards, their resources, any options, consequences of their options, and the relationship between the potential ethical issue and their responsibilities (*particularization*). Finally, they formulate, select, and execute a *judgment*.

To create an ethical mitigation guide, this project relies on the second and third activities: after a worker recognizes and ethical issue and begins the search for an mitigation tool, their goal may be direct (to make a judgment) but in order to select an effective mitigation tool and do so quickly, they must have an accurate understanding of (at least) the features of the ethical problem and the candidate mitigation tools. These details, and the appropriateness of the mitigation tool features for the problem features, are what I'll refer to as "fit" throughout this chapter. Fit likely does not need to be perfect– the engineer will likely have to alter the code some in order to use it on their data or model anyway. This

project investigates worker needs in this regard: what features of fit matter when evaluating options? What features are nice to have, but not essential? What elements of work context matter when seeking fit?

Particularization and judgment are not linear stages– a workers often make developmental judgments, seek more information, and re-evaluate [147]. However, analytically separating information seeking and reflection from the decisions they support will help us understand where framing of interventions should aim to help engineers with building understanding (e.g., informative messaging) or decision-making (e.g., persuasive messaging).

Particularization

In this and other work, I have observed ML engineers as they reflect and seek information about many types of "particulars"– just as in other industries [186], particularization is broad. It can include external information and internal beliefs about features of the circumstance, the stakeholders, the ethical issue(s), options, resources, and consequences. This project directly supports a key activity within particularization: seeking information about options.

Of course, the types of particulars are related to one another. For example, in order to find an option, one must understand relevant aspects of fit between options and problems and identify the features of one's problem and prospective options to see whether they are suited. In order to evaluate fit, one must be able to predict consequences of each option; identify needed and available resources; and define (at least intuitively) success, failure, and acceptable risks.

As part of the empirical investigations, this project observes and accounts for broad particularization among engineers who have recently become aware of an ethical issue, and engineers engaged in the more narrow task of seeking options.

Judgment

Rest's foundational work on what was then called "moral sensitivity" structures my conceptualization of judgment. He identifies three activities: "formulating the morally ideal course of action; deciding what one actually intends to do; or executing and implementing what one intends to do" [148].

This study focuses on identifying ML engineers' intentions, but necessarily remains aware that some may prefer to report what they see as the moral ideal to a researcher; that in a real work situation, organizational and interpersonal factors may influence engineers' intentions; and that their perspective or options could change as they attempt to execute a decision. Therefore, although the designed artifact focuses on presenting options and features of fit, it also supports two secondary goals: education about different conceptions of morally ideal courses of action and tools to facilitate execution in real work settings.

4.3 Design Problem

Technical interventions abound to trying to encourage human values in machine learning (ML): privacy, transparency, and fairness in particular. But how can ML engineers find an intervention to suit a given ethical problem, dataset, ML technique, and domain?

Imagine you are a machine learning engineer and you have recognized a potential ethical issue in your work. Maybe you noticed a performance difference among demographic groups in your model or you suspect it is using a proxy for race, like zip code, to make different predictions for a racial group. You decide to see what you can do about it and search for a popular fairness toolkit that was recommended to you.

On the website, you find links to code, tutorials, a paper, videos, and more. You are interested in determining whether the toolkit can help you diagnose and address the problem

you are worried about. You scroll until you find options for identifying bias in ML algorithms (labeled for example, "Equal Opportunity difference: The difference of true positive rates between the unprivileged and privileged groups;" "Mahalanobis Distance: The average Mahalanobis distance between the samples from the two datasets;" and "Manhattan Distance: the average Manhattan distance between the samples from the two datasets") and options for bias mitigation algorithms (labeled for example, "Reweighing: Use to mitigate bias in training data. Modifies the weights of different training examples" and "Disparate Impact Remover: Use to mitigate bias in training data. Edits feature values to improve group fairness.") Clicking on any of these options brings you to a GitHub page with well-documented code that you can download and start working with right away.

This is undoubtedly a useful resource: it offers all of the features that participants in this study pointed to as desirable: code, tutorials, and videos. However, it may be difficult to navigate without a highly particularized understanding of the circumstance, resources, options, and consequences you are facing. For example, if you are not familiar with what a "Manhattan Distance" is, the fact that you can determine the Manhattan distance between two distributions isn't likely to help you decide whether that is the mitigation you need. Perhaps you notice a technique that claims to improve group fairness, and you get fairly far into implementing it only to realize that it works only for groups that are explicitly defined by a feature in the data or for exactly two groups and no more. The goal of this project is to help machine learning engineers quickly build the understanding necessary to select an appropriate technical intervention in a single, searchable, filterable resource.

This chapter describes a Value Sensitive Design study aimed at developing a guide to ethical mitigations for training data that considers the perspectives and practices of machine learning engineers and supports the interests of lower-power stakeholders– like subjects of training data or citizens who aren't aware decision about them are being made by an algorithm. The values and interests of these data subjects and decision subjects are diverse and

vary not only among individuals and subgroups, but also among ML application domains (e.g., music recommendation, law enforcement, and medicine implicate different values and interests). Therefore, I will instead aim to reduce risks of value violations, interest conflict, and harm to data and decision subjects while making the interface for the guide maximally useful in helping machine learning engineers understand their options and make a selection.

4.4 Design

Value Sensitive Design uses iterative conceptual, technical, and empirical methods to develop designs that reflect the values of key stakeholders. As I worked on conceptual, empirical and technical investigations, I held ongoing conversations with a key informant: a hobbyist machine learning engineer and professional web and mobile phone application developer. This informant was too close to the project to serve as a participant, because he was aware of the study and its larger context, is close to the author personally, and watched the design of this study develop over time. I eventually hired him to build the prototype tool I designed as part of the technical investigations. He was able to offer insight into engineering as a profession and culture, in particularly challenging me to reword and expand on the wording of filter groups, filter options, and fields in the tool profile form so that they would be clearly understood.

This section describes how I used these investigations to "design up" for higher-power actors and to meet the following design goals:

1. Enhance users ability to perceive, particularize, and judge technical mitigations for known ethical problems in training data.

2. Improve awareness of existing and new technical interventions among practitioners and researchers.

3. Empower trainers, educators, and leaders in ML with structured and restructurable information about technical interventions for ethical concerns in training data

4. Achieve above design goals while minimizing interruption to ML engineers' work practices

These design goals were developed during conceptual investigations.

4.4.1 Conceptual

Given that the goal of this project is to "design up," ML engineers, their managers, and educators are the direct stakeholders (they will be using the system). However, the people whose interests are under threat by the phenomenon of unethical ML are the people downstream: users, citizens, data subjects, and underrepresented groups will be affected by the systems the direct stakeholders make. For the purposes of designing up, I considered direct stakeholders needs in terms of usability, and adoptability– optional tools won't be used if they are uninteresting, difficult to navigate, or worse than the existing solution– but chose to prioritize reducing risks to the vulnerable, low-power actors in the system, namely data subjects, citizens, future citizens, and underrepresented groups.

Prioritizing values of lower-power stakeholders in a design for higher-power ones requires a similar "flip" to the ones used by Nader & Barabas et al. [15, 131] This switch is from a defensive posture (encouraging people to protect themselves: read the privacy policy, install an extension, don't use that service) to an offensive one, in which we encourage engineers to catch and deal with potential threats. This guided the selection of ethical sensitivity as a design goal.

To identify further design goals, I conducted a stakeholder analysis. The impacts of machine learning on stakeholders is well-studied problem: I read research articles about potential problems and harms in ML [17, 157, 158], their measurement [8, 25, 33, 37, 54,

109, 124, 171, 182], values and operationalizing them [55, 58, 69, 85, 87, 183], and interventions, their motivations, and impacts [20, 23, 38, 41, 44, 70, 108, 195]. I also included some papers about generalized Artificial Intelligence, after noting that some machine learning is done with the intent of future general intelligence [84, 174]. I identified stakeholders and listed the potential benefits and harms a guide for ML engineers could have for each of them (Table 4.1), listed values implicated by the potential benefits and harms, and identified potential value conflicts among stakeholders. I also retained any paper that described an ethical mitigation strategy in a list so that they could be included in the ML Ethics tool.

Before undertaking the empirical portion of this study, I conducted nine pilot interviews with ML engineers and data scientists aimed at understanding their existing training data workflows. I used this information along with another interview study focusing on the needs of ML engineers in industry [93] to form my initial understanding of their values and interests and to guide the development of the empirical work.

Holstein et al. identified several disconnects between the needs of ML engineers and the offerings of ML fairness research in 2018 [93] that helped me decide to use a series of filters. Holstein et al. identified a lack of tools about data collection (alongside a desire among engineers to intervene in data collection and curation), workers' concerns about their own blind spots about sources of unfairness, needs for proactive and holistic auditing tools, and challenges around addressing problems once they've been detected.

I combined these insights from Holstein et al. [93] with the corpus of mitigation papers I found in my literature review to select the filter groups (objective, development stage, datatype, ethical concern, and ML field) and the options available for each. The corpus of mitigation papers I'd collected made it clear that detecting and mitigating problems can happen at any stage in development, so users should be able to filter both by development stage and objective. Separate filter groups for objective and development stage make it obvious how to quickly find, for example, papers that could help them detect problems while

working with training data, while training, or when working with a completed model. In the filter set ”objective” I started with ”detect” and ”mitigate,” based on the findings from [93]. I added added ”report” based on the prevalence of papers in the corpus (like Datasheets for Datasets [74]) that allowed engineers to communicate about the context, contents, provenance, and ethical issues in their datasets or model. Finally, I added ”plan” to help support engineers as they seek to identify their own blind spots, which [93] indicated was an unmet need, including general papers about fairness definitions and sources of fairness problems. The filter set ”development stage” included collection, training, and post-training, inspired by [70]. The corpus included many papers that were designed for a particular data type (like tabular data) but didn't make that clear in the title or abstract, and also papers published in an ML sub-field that wouldn't be useful outside of it. Furthermore, in the empirical investigations, engineers often added their ML field or datatype as a keyword in their search queries. Filter groups ”datatype” and ”ML Field” were therefore added and filter options were selected based on papers available in the corpus (additional filter options are easy to add as new tools emerge). Finally, this framework of filters gave me the option to expand the scope of the tool beyond ”fairness mitigations” to ”ethical mitigations” without making it less usable for people seeking fairness interventions. To that end, I added the ”ethical concern” filter set and populated it with options represented in the ML mitigations corpus I built during my literature review.

Conceptual Results

Based on the stakeholder analysis, I selected several supported values and wrote working conceptualizations for them, assuming at first that the scope of the tool would be narrowly scoped to fairness mitigations in machine learning.

Usability: the designed artifact should accomplish its other goals with minimal disrup-

tion to existing practices.

Productivity: the artifact should allow workers to accomplish as much or more work with the artifact as they did without it for a similar amount of time or effort.

Fairness: the artifact should support and center fairness-enhancing technologies; the artifact should make it easier for workers to create systems that fit an appropriate definition of fairness [69]; the artifact should allow people (including but not limited to workers) to read, understand, evaluate, and implement several definitions of fairness [70]; the artifact should expose gaps in coverage of fairness-enhancing technologies to encourage their development; the artifact should expose components of ML systems that threaten fairness to encourage their recognition and support efficient particularization in the future.

Adaptability: the artifact's structure and components should be able to be updated as technology and practices change; the artifact's structure and components should be able to be tailored to suit particular situations; the artifact's structure and components should be able to be expanded to encompass other values, new mitigation strategies, and other goals.

While working on the design (as part of technical investigations), I realized that the guide could employ filters, allowing it to easily scale to support other values in ML design without compromising usability. While seeking out mitigation strategies that supported privacy, transparency, and other relevant values, I encountered many useful conceptualizations of those values. As with fairness, the guide allows and supports users as they develop their own conceptualization and does not rely on a single understanding, so the requirements relating to fairness could be reused for privacy, transparency, and other values. I determined the final list of values using the interventions I found and ensured that additional values could be added if needed.

4.4.2 Empirical

I targeted empirical investigations to collect three kinds of data: how participants understood their particularization habits and the impact of their current organizational environment; descriptive data about broad particularization (reflecting and seeking information about many types of particulars); and specific observations of participants seeking options. To those ends, empirical investigations included interview questions about particularization at their workplace, observing engineers while they particularized on their own (with no guide), and asking engineers to search for a mitigation using a tool.

23 machine learning engineers participated in this study. They had between a few months and 15 years experience with machine learning (an average of three years). One volunteer, two ML managers, eight students, and 12 ML workers participated. Two worked or wanted to work primarily in academia, 15 worked or wanted to work primarily in industry, two expressed interested in working in both, and four were unclear or unsure. Participants were recruited from a ML meetup group the author attends (6), referral from other participants (7), and on several internet forums (/r/machinelearning, /r/artificial, /r/datascience, and hackernews.com). Participants needed to be 18 years or older and consider themselves data scientists, machine learning engineers, or people who worked with training data data science or ML algorithms.

Depending on available time and the stage of coincident conceptual and technical work, participants were either asked direct questions, asked to particularize on their own, asked to about a draft in development, or asked to particularize using a popular AI Fairness toolkit available online. Table 4.2 shows how participants were distributed among these activities.

Data collection for this study happened immediately after and with the same participants as another ethical sensitivity study, in which participants were presented with an ethically-fraught and unfamiliar facial recognition dataset and asked to think aloud while

considering using the data. One of the ethical problems in that study (bias in performance by race and gender) was used to frame particularization activities in this study. If the participant particularized that problem on their own during the previous study, that data was considered as well to round out my understanding of unguided particularization.

Questions

Twelve participants were asked direct questions about what information they would look for and from where. Questions included:

> Have you ever encountered an ethical issue in your work? What did you do?
>
> Where would you go for information if you weren't sure about the ethics of something, or to decide what to do?
>
> What sources for information about ethical issues and interventions do you trust?

Particularizing without a tool

Eleven participants were asked to particularize on their own with the following prompt after discussing fairness problems in a facial recognition dataset.

> "For the next step, I'll ask you to imagine that after a few weeks of working with this data, you and your team noticed that there were a lot more men than women and that there were some skin tones not represented well in the data. You're worried that this will hurt the performance of the model for those groups . . . think aloud as you use the internet, your own resources, or reflect on how you might move forward knowing this."

The goals of this task was to collect information about what information, sources, and types participants would search for given unguided access to web resources. If they engaged in reflection, what did they reflect about? What kinds of examples, legitimations, beliefs, and preferences do they rely on when building an understanding of the problem and working toward a judgment? This open-ended task allowed me to collect data about particularization in general among machine learning engineers.

Particularization with draft

The original plan for the empirical investigations was to use them to iteratively develop a draft, meaning that some participants would be given prototypes to use to search for an answer. In practice, however, I found that the low fidelity and non-comprehensive drafts I was able to produce between study sessions were insufficient to generate meaningful data about their utility for particularization. Three participants used a draft until I determined that this problem could be avoided while still getting high quality data about particularization and judgment by instead asking participants to use an existing, thorough, high-fidelity toolkit.

Particularization with toolkit

Six participants were asked to search for a mitigation for a particular fairness problem in facial recognition data using an existing tool kit. The goal of this task was to identify barriers to search, salient or sought for features of mitigation candidates, and types of information they sought about mitigation candidates. This task collected more specific data about machine engineers searching for information about options.

Empirical Results

Participants sought out high-level information sources– like blogs, videos, and Wikipedia articles– along with academic articles and code. They wanted to know how candidate interventions worked and how they fit with the problem at hand. Participants discussed seven aspects of fit, five of which are supported in the final design.

The following section provides detail on the sources and types of information participants sought, their reasoning, how the information they found contributed to particularization. It also identifies the aspects of fit participants were interested in and explains how I selected 5 to support in the designed guide.

Information seeking: sources and types

Participants relied on secondary sources, like blog posts, videos, and Wikipedia articles, for general guidance and primary sources, like code and academic articles, for detailed understanding when seeking information about ways to mitigate fairness threats in facial recognition.

High-level sources Participants used blog posts, videos, Wikipedia, and similar summaries of techniques, problems, and interventions. They tended to use these either as a primer to understand how a technology works or to refer them to more specific resources.

This study recruited people from all areas and levels of experience with ML to do a task in a fairly well-defined area (facial recognition) so necessarily, participants' experience with the task varied. Participants who had less experience with facial recognition or computer vision in general used high-level secondary sources to understand the technology. For example, P8 searched "how does facial recognition work?" during particularization without a draft and found a video on YouTube by the same name. They watched parts of it, then scrubbed through, looking for information about how images are processed. They also searched Google for "face detection" and selected Wikipedia. After building more

technical understanding, they searched Google for ”bias in machine learning facial recognition,” selected a medium post, followed a link to ”Man is to Programmer as woman is to homemaker? Debiasing word embeddings.” [25]

P19, P8, and P20 used or mentioned Medium.com, and P7, P8, and P20 mentioned or used towardsdatascience.com as useful high level sources. P20 explained their use of blogs in response to interview questions:

> “So you know, when I’m looking into Towards Data Science, or Medium or any of these other blogs, I’m looking for things like this, you know, the resources that they are pulling from so I can go direct to– OK perfect. One click and I’m already at a potentially good, you know, article, research paper, etc.” -P20

Some participants preferred to orient themselves to a mitigation strategy they were considering using more high-level content, like introduction text and videos. While particularizing with an existing tool, P23 explains:

> “The first thing I look for is like a brief intro. And, you know, demonstration or what a brief introduction on what, what each algorithm can do, and in which situations it can be helpful. So that’s the first thing, the most practical thing. And I see now that there’s some videos here, I’ll probably look into this as well. But my first big reaction is to get as much information, practical information as I can. So my first the first thing I want to see is what types of our teams they have. They have and then What do they do? And then how do they work and then have to see the code eventually, but I will first get the general sense.”

Primary Sources Participants used two types of primary sources: code and academic papers.

While reviewing a draft, P6 explained that high-level content is useful, but under time pressure they would go straight to the code.

> "If I'm just kind of working on something like leisurely, I'll watch the video and see what's up and maybe read a little bit about it. But trying like, hey, we've got this arbitrary deadline . . . I'll get the code working. And then in the process of getting it working, that's when I'll actually learn, like, everything it's doing, which is a little bit easier than reading the whole thing, then putting it in and trying to get it working on, it saves a little bit of time."

P8, P15, P16, and P23 mentioned looking at code as well.

P4, P15, P18, and P20 used or mentioned academic papers or posters. P1, P4, P5, P6, and P10 discussed the credibility of academic papers and referred to them as a useful source.

Participants who discussed academic papers often engaged with questions of credibility.

In response to interview questions, P1 said "I tend to trust Google and all the academic papers that they produce." They acknowledged that Google has struggled with ethical issues of their own, but that "think there's a lot of tools that they provide that give you analytics, in terms of geography, of demographics of people and those sorts of things." Discussing a draft of the guide, P5 wondered aloud about the credibility of papers on arxiv, a popular source of pre-prints, white papers, and unreviewed computer science papers: "would that be high enough quality? . . . sometimes you want to see what other people are doing, but it's sort of not up to par necessarily with [peer-reviewed] publication."

Searching without a tool, P4 used a fairly general search term ("facial recognition machine learning") to find a secondary source from machinelearningmastery.com. They then followed a link in that web page to find an academic paper written by Viola and Jones called "Rapid object detection using a boosted cascade of simple features" [180]. They used Wikipedia to contextualize it, returning to machinelearningmastery.com to follow a link to the Wikipedia page for the Viola-Jones detection framework. On the Wikipedia

page, they highlighted text in the "Learning Algorithm" section. They eventually returned to the academic paper, scrolling first to the "Features" section and stopping to read. They continued to scroll, looking for text that would help them understand the functions:"So yeah, I'm checking that here. They're using the h_t, and I was trying to figure out what h_t was." When discussing the purpose of reading the paper, P4 said "So what I would do, like once I read the relevant foundational papers, maybe I would try to implement whatever they did, there might be some packages . . . And then if I see like some parameters that might be modified to better suit my, my task then I would try to modify it, or if it works well and within the limits– within some acceptable limits, then I would go ahead and just use that."

Participants appeared to trust academic papers, but often relied on summaries of papers on Medium.com, towardsdatascience.com, and other sources to ensure the relevance of a paper before downloading it. While reviewing a prototype, P6 indicated a preference for summarized content: "I kind of get annoyed with when I'm looking stuff up. A lot of the time you have to, to find exactly what you're looking for, you have to scroll through a whole paper. Whereas with this, yeah, with this, with how this [early prototype] is set up, you can find what you're looking for, and then read through the paper, which is ideally the way you want to do it before you waste your time reading the whole paper about something."

In response to these findings, tool profiles operate like a high-level resource (explaining the purpose, requirements, and operation of each strategy) but also consistently and clearly link to detailed sources, like papers and code.

Information need: The "How"

Regardless of what sources they sought or terms they used, when considering a candidate mitigation strategy, participants wanted to understand how the mitigation works: what they would need in order to use the mitigation strategy and what, specifically, does it *do*. The design I landed on included linking to code, papers, and "other links," which may be tutorials, videos, and demonstrations as features of the tool profile. But how could I sur-

face the "how?" While answering questions about particularization habits, P7 offered the metaphor that inspired the final design for surfacing this essential information:

> "But when it comes to like, time constraints and you really are trying to extract some useful information out of it, then I would just like, go to the important point pointers, like what are the ingredients and what is the procedure? And so, because that's the first thing I would obviously look at as like, ingredients, if if I have the ingredients only then I can move on to procedure because there's no point of doing the procedure and when you come back, like come to the Step five, you realize that there are no ingredients." *-P7*

"Ingredients" and "procedure" became fields in the initial design, and are now called "requirements for use" and "overview of procedure," after my informant encouraged me to clarify.

Information Need: Fit

Many features of an ethical mitigation can cause it to fit or not fit a given ethical problem: this study identified seven aspects, five of which are supported in the final design and two of which are not supported. Figure 4.1 abstracts fit, work engineers have to do to make an intervention better fit their problem, and gaps that the mitigation doesn't address.

Many participants' search terms included more than one feature of fit, for example: "bias in machine learning facial recognition" (P20), reflecting the need to filter results by multiple areas of fit– a ethical issue (bias) and an ML field (facial recognition) at the same time.

While looking for a mitigation with no guide, P6 discusses the use of high-level and specific sources toward their goal: "what I usually try to do in this situation, if I am looking at just random code and stuff, instead of reading all of this on through, the first thing I'll try to do – he's attached a video here, so I might just watch that and see how it works instead

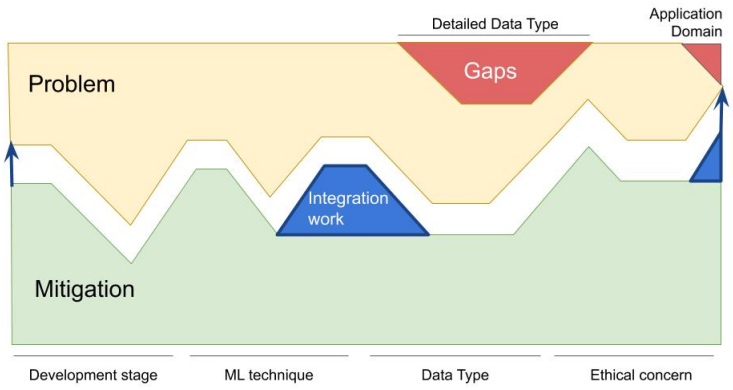


Figure 4.1: Fit between problem and mitigation requires integration work and the mitigation still may not cover gaps

of having to read it all, but I usually go to just the download or just copy and paste all of this [code] right here. And basically the first thing I want to do is get it running and see how it works, what it does, and what kind of changes I have to make to it to make it useful for me."

Although P15 would have preferred to find a mitigation that was tailored to fit the facial recognition problem, while particularizing with an existing tool, they found paper describing a fairness intervention for a binary classification problem. "the issue is that if it's only say defined for binary classification, then it's not really that relevant unless we can formulate our problem in such a manner," but also said "I'd certainly keep it in the back of my mind."

These quotes illustrate a pattern: participants wanted to know how much integration work they had to do– in other words, what changes would they have to make to the strategy as proposed to make it fit their problem. Generally, participants aimed to find a mitigation that requires less integration work, rather than more, however, as P6 implied, there is always *some* integration work that needs to be done, so a the lack of a perfect fit is not a deal-breaker.

The design supports four areas of fit using filters: development stage, ML technique, data type (broadly), and ethical concern. Filters (see Figure 4.4) allow searchers to find mitigations that fit all four, if they exist, or to broaden their result set by prioritizing a subset. It addresses less common areas of fit, coding language or environment, and other specific requirements in the tool profile (see Figures 4.2 and 4.3).

Development Stage Development stage is a common way of classifying ethical mitigations in ML into three categories: do you intervene in the input, the process, or the output [70]? Participants rarely included developmental stage terms (like "training data" or "before training") in their search terms, but it was frequently a part of participants' problem framing: they considered image augmentation and manipulation techniques rather than training or post-processing mitigations.

P18 brought up the need for more exploratory tools while particularizing with an existing tool and, in doing so, revealed an awareness of the need for developmental stage fit.

> "Yeah, I think it's important to do more detection because this [NeurIPs paper describing an intervention] is more about post-process bias mitigation. I'm not sure how you could just choose this out of the box on your data . . . I think this will definitely require someone to actually know what the data is."

This comment about post-processing, though, came after seven minutes of searching through the paper and discussing whether it actually intervened post-processing, or whether it detected problems post-processing, but intervened in training data. This is an easy confusion to address in a search tool. I tried to clarify by using "collecting/cleaning data," "training by model," and "post-training" in order highlight developmental stage as an important problem feature and improve clarity for users. Comments like P18's also indicated a need for the "Objective" filter.

Objective As P18 mentioned, identifying and measuring the extend of an ethical concern can offer important information to guide the selection of a mitigation. These detection tools can also support an engineer who needs to approach their client or manager to say, " we need to spend some time or resources to address this."

These two objectives, "detect" and "mitigate" were the focus of all the options-seeking search and browsing behavior among my participants. However, two other objectives were represented in the list of interventions I compiled during the conceptual investigations, and which I thought were important to include: planning and reporting. Papers that may help with planning include reviews of interventions in an area (e.g.,[55], papers that disambiguate important concepts and present formal models (e.g., [58]), warn of unanticipated problems (e.g., [183]), or propose a new high-level approach (e.g., [109]). Reporting papers generally offer standard documentation that can be used to describe a data set (e.g., [74]) or model (e.g [127]). Planning and reporting resources may benefit ML engineers, but engineers may not be aware of them. I hope that including them as filter options and tags on tool profiles will raise awareness of their existence.

ML Field or Technique Many participants included ML field in their search queries, cited it as a reason for looking further into or disqualifying a mitigation strategy, and relied on it when contrasting the task at hand with their own experience. It also stand to reason that an engineer seeking a paper about working with word embeddings need not scroll by papers about facial recognition and vice versa.

However, selecting the options for the filter categories proved to be more complicated than I expected. Some fields, like Natural Language Processing, seemed to be fairly well defined, both in participants' discussions and in the literature. But others were not as clear. Should facial detection and object detection be in the same category, or separate? For example, responding to an interview question, P19 noted similarity between their work and the facial recognition task, but ultimately evaluated them as different:"For me personally,

this is a new domain. a lot of the machine learning I have experience working with . . . they usually us biomedical or biology examples . . . still computer vision, but a different type of image."

Ultimately, I decided use the list of mitigation strategies I'd already collected to select filter options: if there were more that two mitigations in the list designed for the category, I included it. This meant that face detection recognition did get a category separate from other types of computer vision. To ensure that this initial decision doesn't limit the usefulness of the tool if things change, the final intervention uses a filtering system that makes it fairly easy to add new categories.

Data Types When discussing the potential usefulness of this tool, participants emphasized data type in their searches and discussions.

P5 mentioned the most specific search patterns of any participant while we talked about their workplace after searching with a draft. "I would specifically, like look up MRI data. I would look up whatever I want to do 'for MRI data' . . . First by the sort of body part, and then modality." When discussing the prototype for a search guide, they mentioned they'd be interested in such a feature, "So maybe sorting by your data set would be more helpful than by algorithm . . . then if you can think of like different sorts of data sets and try to find out where the bias comes from in each." When discussing the tool prototype, they suggested searching by data type to allow reasoning about causes of bias: "So that would be interesting to look at too– what the bias be in that data set and then if you can think of like different sorts of data sets and try and find out where the bias comes from in each"

Data type is included as a filter with high-level options derived from the mitigations list compiled in the conceptual investigation: text data, image data, tabular data, and other data. Granular data types, like brain scans or X-Rays, can be included in the tool profile, but are not included as filters: there are so many options for each that it would overwhelm the user interface and many of the options would likely return few or no mitigation profiles when

combined with even one other filter. Also, a tool developed to ensure an important value should not be hidden from the user because it was developed for chest X-rays if it could be adapted to work for other types of diagnostic images. Users can see this information in the tool profile and can use the search feature to find it, but search terms are combined using the Boolean ”OR” rather than ”AND” to increase the number of potentially relevant results.

Coding Language While particularizing with an existing tool, P18 mentioned that they prefer finding solutions that are built for the coding environment or language they work in: “But I work in MATLAB so I always use something from MATLAB . . . So I always do it like manually or something, if it already exists, in MATLAB.” also indicated that this is a guiding feature in search: “Specifically for myself, I work a lot with Python. So [search terms] as simple as ‘Python, open source, ml’ tends to really narrow down the topics that I’m working with.”

I included a spot in the tool profile to provide coding language, but did not include it as a filter. Few mitigations in the compiled list included code; most can be implemented in any coding language. Right now, a coding language filter would cause mostly blank queries, which may discourage searchers. Therefore, I included a space in the tool profile for “languages supported,” to ensure that any mitigations that do use code have their languages represented in the profile and so that people use the search function in concert with filters to see whether an intervention for their problem using their preferred language exists.

Unsupported fit There are two related features of a problem that emerged in the study, but I decided were too granular to create filters for: application domain and detailed data types. These were both discussed by P5, who works works with medical imaging data and indicated that they would search for mitigations and include body part and imaging modality (e.g., “MRI” or “X-Ray”). The tool does not allow for filtering by application domain

(e.g., medical diagnostics) or detailed data types (e.g. MRI data) because the number of filter options would be very high and I encountered very few mitigations in my search that were so narrowly scoped. This would mean that most combinations of even one other filter (e.g., a search for "data type: X-Ray" plus "development stage: training") would be likely to return zero results, frustrating users. As a compromise, the tool does include a (rudimentary, for now) search feature, which would allow a user to select the filters they want and the application domain or specific data type in the search bar.

ML engineers will need to do some integration work and may need to fill some gaps on their own even after finding a useful mitigation strategy (see Figure 4.1, but the tool developed in this study will allow them to more quickly identify mitigations that fit key problem features.

Persuasion

The stated goal of this design project is to help ML engineers understand ethical problems in their work and select appropriate mitigations. "Studying up" and trying to understand the work lives and perspectives of engineers rendered another need this tool can help ML engineers fill at work: persuasion. Participants discussed communication barriers that make necessary communication with clients and decision-makers difficult.

While discussing a draft, P5 talked about the difficulty engineers face when trying to advocate for ethical issues.

> "But there's this problem where even if engineers explain everything right, like have all the facts, know the theory, try and explain the theory and the most common like layman terms ... If they have an idea you can't really convince them ... higher up people are like we have a deadline to meet, we can't do it."

P5 used the Challenger explosion as an example, concluding: "Engineers just had to do it at that point. It's like, you're a cog in the machine. If you don't do it, they're going to

find someone else to do it for you."

While particularizing without a guide, P20 also cited difficulties communicating with clients and people in their organization. When explaining why they selected an academic article when looking for an ethical mitigation, P20 explained:n

> "Again, this is peer reviewed. You know, some of that is important. Some of it– in the business sense really isn't: if it works, it works. I don't really need to know all the sources necessarily. But something beyond 'I've done a Kaggle[1] exercise.' It's kind of nice for a– if I need to tell my bosses why I spent three weeks on something that came up blank. It's nice to not say 'hey, a junior in high school, wrote a Kaggle post on it. I thought it looked great like that.' That is nice to have kind of some backing as to like, 'Hey, this is the research that was going off of.' "

Participants were interested in making ML that didn't violate their ethical standards or hurt others. Some used a profit-motivated legitimation (i.e. a failure would be disastrous for public relations) and many used moral legitimations instead or as well. However, when imagining discussing the problem of fairness in facial recognition during the study, they seemed most comfortable with a quality framing, rather than a moral one.

After looking for a mitigation using an existing guide, P9 gave the example of an algorithm trying to diagnosing X-Ray images sourced from two cities: one with a low positive rate and another with a high one.

> "At first they they trained it and the the stupid algorithm immediately was like, Oh, well, the, the resolution is different on the two, or the the file format . . . then they change it. And then like all the resolutions different and so they like change them . . . And then eventually, when they thought they had scrubbed

[1] https://www.kaggle.com/

> everything out, then it was like the little patient ID number tags were different shape. And so it's like, well, the rectangular ones are from New York and the like rounded corner ones are from Ohio . . . They kept trying to remove all these things and the algorithm predicting higher rates for images in New York, just because it knew that it would have been trained on a higher prevalence rate in New York . . . it was just too good at it."

Referring to that example, P9 said, "Actually, I really like that one because I think it's really instructive. It doesn't have all the charge about like racial bias, gender bias . . . At least like we can all agree without getting into the thick of the politics, right? We don't want false negatives in Ohio."

Other participants prefaced their discussion of ethical issues in facial recognition with phrases like, "I don't know, we– I don't want to get too much into politics here, I guess" (P3).

Engineers indicated their need to justify their technical choices and time to clients and managers. They tended to believe that this communication was better supported by quality and profit legitimations than moral ones. The guide may in part be useful to engineers by providing credible support and motivating examples when discussing ethical problems and potential solutions with others in their organizations.

4.4.3 Technical

The technical investigation integrated findings from empirical and conceptual investigations into an filterable guide to ethical mitigation strategies. The literature review in the conceptual investigations revealed that engineers want detection and mitigation tools, but also auditing tools and ethically-aware data collection support [93]. Through empirical investigations, I identified some key features of ML engineers' particularization that

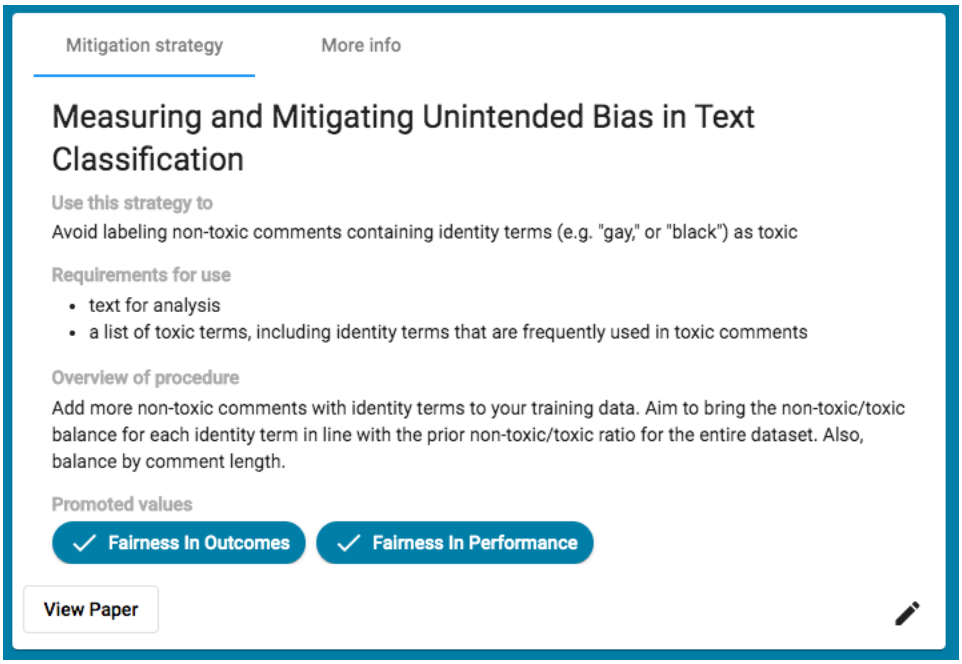


Figure 4.2: First Page of Tool Profile

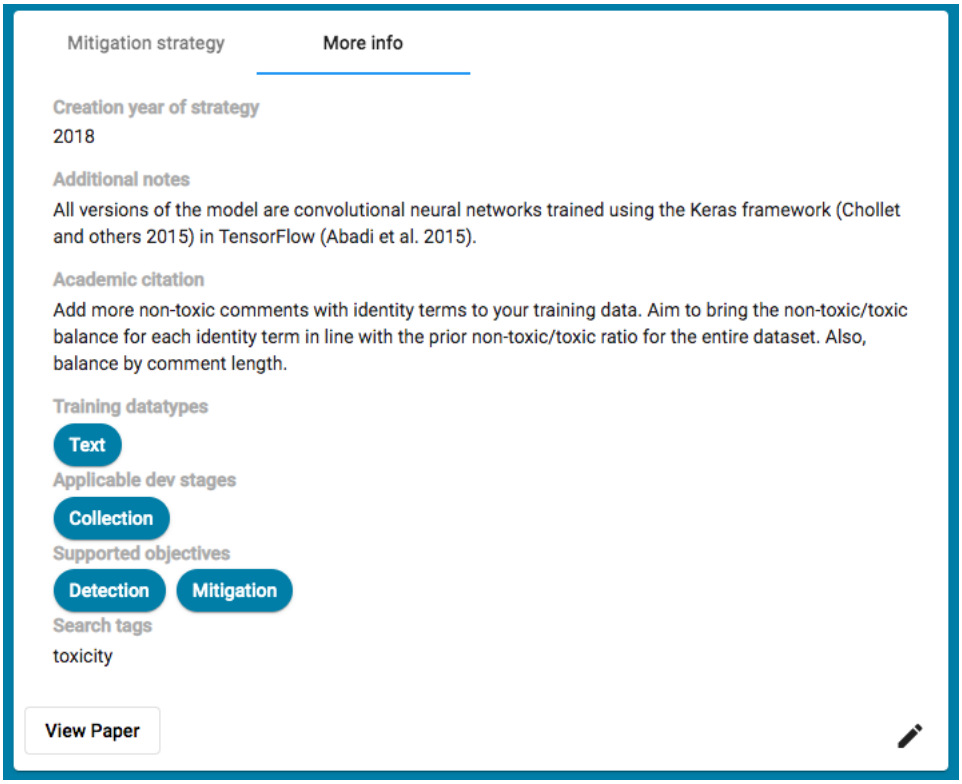


Figure 4.3: Second Page of Tool Profile

informed the technical investigations and the design. Participants demonstrated interest in and discussed several aspects of fit, including objective, ethical problems, developmental stage, data type, and ML techniques. They were interested in code, papers, and videos or other tutorials. These needs are reflected in the design as features of the profile for each tool (see Figure 4.2 and Figure 4.3) and sometimes as filters users can select for search (see Figure 4.4).

Instead of making an inexpert prototype myself, I opted to hire a professional programmer. This programmer was not a participant in the empirical investigations, but is both a web developer and an ML engineer. He served as an informant for this study from the beginning and offered valuable technical and usability insight to the prototype. Notably, he helped me phrase my filter categories more clearly. For example, I had mitigation strategies classified by ”group” (including “detect,” “mitigate,” “plan,” and “ report”) and ethical issues (labeled as “fairness: outcomes,” “fairness: performance,” “privacy: sensitivity,” “privacy: security,” “accountability: transparency,” and “accountability: interpretability”). He encouraged me to expand on category names and we rewrote the filters as phrases, like “My objective is to: detect ethical issues in my model,” or ”mitigate an existing harm.” ”My ethical concern is: reducing unjust discriminatory outcomes,” or “ensuring equal performance across subsets.” The result of hiring this programmer is a prototype that is competently built, an interface that is legible to people in the community, and code that is comprehensible to people who would like to contribute to the open source project.

Evaluation

To help evaluate and improve the tool, I reached out to ML engineers (some of whom were in the study and some who were not) and technology ethics researchers. Two of each offered comments.

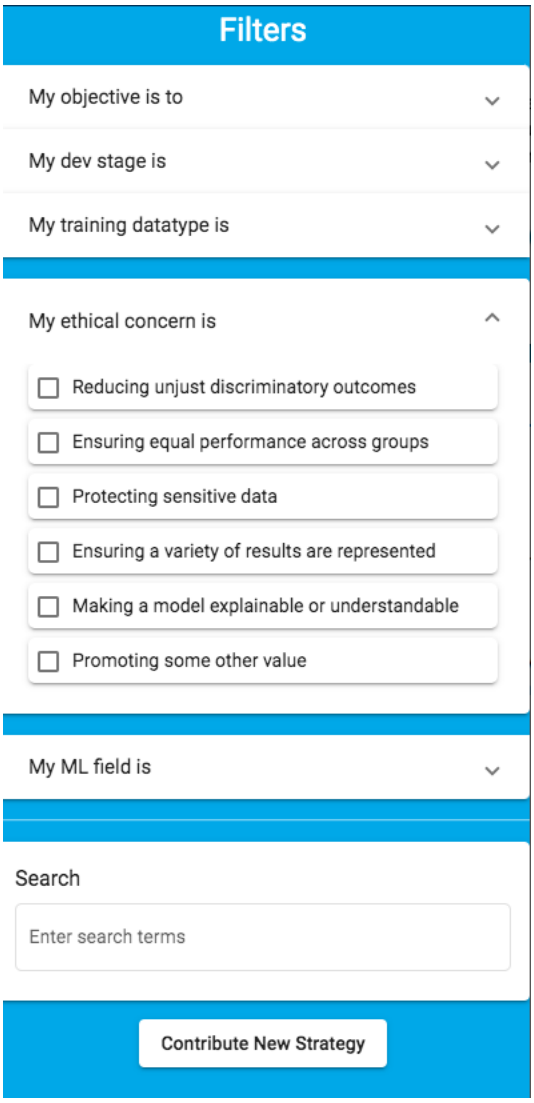


Figure 4.4: Filters, featuring options for ethical concerns

Four suggestions were implemented based on their feedback:

- **Users should see radio buttons (instead of check boxes) when selecting filters.** Radio buttons communicate that one option can be selected at a time, where check boxes suggest that users have the option to select more than one option. Making this changed represents a trade-off: radio buttons risk implying that one option should be chosen for every filter set, which I consider to be limiting: I hope users select as few filters as are useful to maximize results returned and encourage adapting interventions across circumstances that are not exactly the same. However, feedback made it clear that confusion generated by check boxes outweighed this concern.

- **Users should be able to see how many results are returned with each search.** "Displaying [number of results] results" at the top of the tool profiles list helps users orient themselves and understand how many options they have. It also may help researchers or other people trying to see landscape of options and gaps.

- **Users should be able to see which filters are active when filter sets are collapsed**. The "Active filters" box was added so that users can see which filters are acting on the result set.

- **Users should be able to clear all filters with one click.** Users can now select "clear all filters."

Known Issues One of the technology ethics researchers who reviewed the tool pointed me to an article urging AI ethicists to consider the cultural and regional context when designing guidelines for AI. I am considering implementing a geographic indicator, but need to carefully consider whether the regions or countries of the publication, author(s) origin, or author(s) institutions should be considered; how to deal with multi-authored papers; and

how to present and enable useful search for this information. I will seek more feedback and consider the above questions further before implementing this feature.

For now, search is rudimentary: if you enter a single word as a search term, exact matches will be returned (e.g., "race" will return results with "race" in them, but not "racial" or "ethnic.") This is the most pressing issue and was noticed by evaluators: users are accustomed to very responsive search features. However, they are costly. I welcome contributions through github to improve the search feature, otherwise, the search feature will headline applications for grants to fund improvements.

4.4.4 Future Development

A technology ethics researcher noted the potential for the tool to be adapted to support group work. I can imagine the interface allowing participants to bookmark and share tool profiles with one another or to collaborate on project-specific lists. I don't know enough yet about how teams work on ML projects or how particularization and judgment play out in groups, but I am excited about the prospect of expanding or tailoring the ML ethics tool to support group work.

I will continue to seek feedback on the ML ethics tool. Because I am particularly interested in including the perspectives of people engaged in ethical technology development, I will seek feedback at a CSCW "Beyond Checklist Approaches to Ethics in Design" workshop this year. I will solicit and monitor contributions through GitHub [2].

New interventions for ethical machine learning that are within scope of this tool are released often. I am aware of several additional interventions that need to be added, and I am sure there are more, especially if academic papers are not published about them or if they are published in venues I am not aware of. I am eager to welcome others interested in

[2] https://github.com/bsmith418/ml-ethics-tool

machine learning ethics, builders of tools, and students to help expand the list of included tools. Any user can submit a tool profile through the "Contribute New Strategy" feature, where the profile draft will go to an administrator (me, for now) for approval to ensure quality.

4.5 Conclusion

This project won't solve ML bias.

First, and most importantly, ML engineers are not causing bias. The training data they are using to build their models reflect the faults of the social systems that generated them. ML development represents a good opportunity to intervene because it is somewhat upstream and can be seen as sitting at the mouth of a new branch. Addressing bias in ML systems to assess risk for parole won't fix the diffuse upstream causes of bias in legislation, law enforcement, courts, prisons, employment, health care, housing and more, but it can prevent those various bias types from being propagated and reified by yet another system: a particularly impactful, opaque, and difficult to change one. In order to pursue justice effectively, society must identify and address the diffuse upstream sources of bias.

Fortunately, addressing bias in the design ML systems does not hinder the effort to attack bias at the source, nor will success at a better, larger justice movement render this one a waste. It's my position that, just like studying up, designing up is a "yes, and" project: in order to catch and address ethical problems in a complex and dynamic social system, we need actors all along any given design pipeline to be engaged. We need to design plug-ins for users that block cookies; we need managers, engineers, and educators who are committed to aligned technology; we need regulators who understand the technology and are interested in disrupting harms its use can cause; we need the companies who invest in potentially harmful technology to be motivated to protect human values; and we need

researchers to rigorously observe the data ecosystem that fuels ML technology.

So we won't solve bias with a field guide. What did we accomplish?

We now have a search tool for ML ethics strategies. Anyone can add to it, and when it is released open source, anyone can expand, tailor, or re-purpose it. In particular, I would like to add a glossary that gives definitions for terms users may run across in papers that may be confusing, like "statistical parity," "demographic parity," "equality of odds," and "equality of opportunity." As it stands, users do not need to be immersed in the ethical algorithms research discourse to understand and select a mitigation strategy using this tool, however, the better they can parse the papers and the more concepts they are familiar with, the more efficiently they will be able to understand and implement them. Also, if a user looks up "differential privacy" they can encounter other ways of operationalizing privacy, further expanding their inner toolkit.

We also learned some things about ML Engineers' views and particularization habits that can serve us in the future.

Most participants in this study were fluent in moral evaluation and legitimation, but sometimes hesitated to use it. Perhaps, when it doesn't compromise the purpose of your project, try to find a quality framing to discuss your concerns, rather than a moral one.

ML engineers are often most interested in the "how" of an intervention. Marking this information clearly or surfacing it in an interface will make it easier for ML engineers to use your tool. Further, if you want an ML engineer to see something, considering placing it somewhere in the path to "how" or in your code, tutorial, or video you made to explain it.

Finally, findings from this design project have implications for the product managers and executives that oversee ML development. Engineers need to feel and be empowered to bring up potential ethical issues. Managers should work to convince engineers that they will not be replaced or punished if they express ethical concerns, but rather that their

technical knowledge and ethical perceptions are valued. Give them resources about, even training for, and time to implement ethical mitigations. Firms are welcome and encouraged to use the ML ethics tool designed here, or fork the project and develop one tailored to their domain. ML engineers are uniquely positioned to notice, understand, and prevent potential downstream harms from the technology they build. Let them.

Table 4.1: Stakeholder Interests

Stakeholders	**Benefits**	**Harms**
ML Engineers	- could reduce time cost of intervening - could improve system quality - could reduce risk of catastrophic reputational harm	- could expand job responsibilities - could compromise overall/average system performance
Managers	- could reduce the risk of catastrophic reputational harm - offers new, free training tool for new workers	- could increase the time cost (duration and frequency) of engineers intervening - could extend time to market
Company	- could reduce the risk of catastrophic reputational harm - reputational benefits from visible social responsibility	- extend time to market - could be seen as taking responsibility for ethical consequences of products
Educators	- offers a new, free training tool for new workers - is flexible, extensible, alterable	
Researchers	- could broaden exposure for ethical algorithms research - could increase adoption of ethical algorithms findings	- could draw unwanted, critical attention from people who are politically opposed to ethical algorithms research
Data subjects	- could increase consideration of data subjects' interests in development	- could increase the need to collect sensitive data - could compromise overall/average system performance
Decision subjects	- could increase consideration of citizens' interests in development - could decrease accumulated/compounded harm over time	- could increase the need to collect sensitive data - could compromise overall/average system performance
Under-represented Groups	- could increase consideration of URGs' interests in development - could reduce discriminatory decisions or disproportionately poor performance	- could increase the need to collect sensitive data
System users	- could reduce risk of catastrophic reputational harm	- could reduce overall/average performance -could reduce confidence in system, decisions
Governments	- could reduce inequality, harms to citizens	- could reduce industry productivity - could reduce global competitiveness

Table 4.2: Empirical Investigations and Participants

Exercise	**Participants**
Questions	P1, P2, P3, P6, P7, P8, P9, P11, P12, P14, P20, P22
Particularize without tool	P1, P2, P3, P4, P7, P8, P9, P10, P12, P13, P19
Review or particularize with draft	P5, P6, P9
Particularize with existing toolkit	P15, P16, P18, P19, P21, P23

Chapter 5: Conclusion

> *"Your scientists were so preoccupied with whether or not they could, they didn't stop to think if they should."*

If you've spent much time reading and talking about ethical technology development, you've likely heard a discussion start or end with this famous sentiment expressed by the fictional Dr. Ian Malcolm[1]. Referring to "Jurassic Park" offers some comic relief and the quote makes a useful point: some technology (perhaps cloning dinosaurs, or hiring by algorithm) is too harmful or dangerous to build. Surely I don't have any problem with encouraging builders of technology to critically reflect on their design decisions or the project as a whole or with supporting refusal to build dangerous technologies. However, I do think the question of responsibility for harmful technology shouldn't stop with implying that its builders are responsible.

Just as the individual right to refuse to use a technology is not enough [22], neither is the ability for an individual engineer to refuse to build. Engineers are individual people acting in a social system much larger than themselves. We saw in Chapter 4 that engineers may not feel able to speak up about ethical issues for fear of losing their jobs.

Perhaps the people building the technology aren't the only ones "preoccupied" here: perhaps the firms themselves are distracted by shareholder value? What are managers' responsibilities? What about governments, who fret and posture about their country's com-

[1] Jeff Goldblum's character Ian Malcolm in the film "Jurassic Park" (1993). An recent example of its use in a technology ethics article is here: [82]

petitive position in AI development, inspiring commentary about a figurative [142] and literal [4] arms race? How about the powerful people and institutions that cause employment, law enforcement, or financial data to be biased in the first place? In general, I support intervention throughout the ML ecosystem, and made my argument for intervening with ML engineers in Chapters 1 and 4. Given that engineers can help with the larger project of ML ethics, it doesn't seem responsible to take a fictional character's word for what builders are "preoccupied" by: what distracts or disrupts engineers from noticing ethical issues, understanding them, and taking action? Perhaps we can prompt and support engineers as they "stop and think if they should." How?

This dissertation uses ethical sensitivity to describe the process by which engineers grapple with ethical issues in training data and to suggest and test design interventions to prompt recognition, support particularization, and enable judgment. It recognizes that engineers are not solely responsible for ML ethics, but sees them as powerful and important people to engage in the project of mitigating harm that can be caused by ML.

5.1 Guidance for Intervention

Far from single-minded obsessiveness with technical details, the machine learning engineers I studied recognized, particularized, and made ethical judgments while working with unfamiliar training data. They expressed interest in building ML that they saw as morally neutral or doing good in the world and avoiding ML that generates harm. They took some responsibility for the ethical implications of the technology they helped build, but recognized the need to and difficulty of convincing managers and others in their organizations. They demonstrated their ability and willingness to think critically abut their work, its social context, and the social impact of technology.

I was surprised and encouraged by how many participants noticed and mentioned eth-

ical issues during a research study that was explained only as interested in "work practices," and by how quickly and confidently engineers identified ethical issues when asked. It suggests to me that engineers have strong ethical senses and skills and that the more an engineer believes ethical sensitivity is valued, they more likely they are to speak their concerns. Engineers' technical knowledge makes those ethical skills keen, actionable, and vital: this section will lay out how I believe researchers, managers, and others can intervene to support and leverage these ethical skills.

5.1.1 Intervention Design

Engineers are not allergic to talking about ethics, but framing outcome differences, performance gaps, opacity of decision-making, and other issues as problems of quality rather than ethics may be more strategic in some cases, allowing them to stay within the bounds of their jobs as they understand them and giving them a legitimation they are more confident using with others in their occupation and organization.

The think aloud study suggests that encouraging ML engineers to consider the characteristics, context, and social aspects of data and technology often results in ethical engagement. Education, training, and workplace prompts to critically consider the data sources and structure and the impact of technology may give engineers some ownership over ethical engagement in their work. Context documents, like Datasheets, offer a flexible, promising prompt to critically consider the social context of ML models and data, which this thesis demonstrates can prompt engineers to "consider if they should."

Context documents like Datasheets do not require capital outlay to use, but they do require a significant amount of engineering time (or the time of other sufficiently expert workers involved in data curation) to write. This study can't quantify the marginal recognition frequency that implementing Datasheets would bring or the cost, but it does suggest

that engineers tend read them, tend to mention ethical issues more often when they read them and that, when they are available, engineers rely on them to build and understanding of the data. Writing these context documents has a significant upfront time cost, but the can be reused and updated for the life of the data set. They can offer important information that workers may otherwise have to guess or spend their work time seeking out from the dataset curators. In this way, Datasheets can support more effective and efficient particularization. Each dataset curation team must come to their own conclusion about whether writing a Datasheet is worth it, but these studies suggest that they support ethical sensitivity and save some engineering time over the life of a often-reused dataset. I hope that many teams and firms will use this evidence as encouragement to implement Datasheets and similar interventions.

The ML Ethics Tool developed in Chapter 4 offers another such tool that can be used as is, expanded, or tailored to a specific work environment. The thesis also offers a conceptual and operational framework that can used or built on to develop (Chapter 4) and test (Chapter 3) other tools designed for the *use* of ML engineers and the *protection* of downstream, lower-power actors.

Finally, any intervention into ethical technology development requires institutional support. Engineers in these studies mentioned feeling unheard by decision-makers or seeing themselves as a replaceable "cog in the machine" at their jobs. Several mentioned wanting a third-party board, agency, or watchdog that an engineer could get advice from or report their concerns to. And a document, training exercise, or code of ethics can only be ineffective if engineers believe they will be ignored, penalized, or fired if they follow its guidance.

5.1.2 Scoping the task(s) of engineering

I suggest intervening in the structures, norms, and policies that inform engineers' understanding of their work, its relationship to ethics, and institutional reaction to their ethical engagement. These interventions can help engineers recognize and raise ethical issues by convincing them they will not be replaced or punished; offer high-quality resources to support particularization; and protect time for necessary technical work to mitigate ethical threats. If you want to engage ML engineers in the project of ML ethics, it will be much easier for them to speak up and to act if they understand ethics to be part of their job responsibilities.

Organizational reforms (e.g., altering job descriptions, training, evaluation and promotion criteria, and resource allocation) need to contend with engineers' perceptions about the scope and precarity of their jobs. Clarifying to ML engineers, managers, marketers, and legal counsel what engineers' responsibilities and abilities are will allow organizations to more fully use ML engineers' skills and expertise. It will also empower ML engineers to act when they need to. or know where to get institutional support for decision-making and execution.

To ensure that ML engineers and others in the organization feel comfortable engaging with ethics in their work, the first step is to ensure that they will not be punished or replaced for reporting or working to mitigate an ethical issue. Then, a firm can work to convey that ethical sensitivity is not only not punished, but valued among employees. Organizations can credibly signal that ethics is an important part of engineers' work by incorporating ethics and ethical work in job-defining practices. For example, ethical engagement can be evaluated alongside other important job aspects in annual reviews. Similarly, training, team-building, software infrastructure, workflows, and human resources processes can include the practice, development, and support of ethical skill. Perhaps a task like the one

in Chapter 3 can be employed as part of a job interview. This would serve a dual purpose: clearly communicating to new hires the ethical priorities of the company and screening out applicants with poorly developed ethical sensitivity skills. If large or many companies adopt ethical sensitivity as a hiring priority, it may encourage college professors, instructors of online courses, developers of resources for ML education, and prospective ML engineers who take responsibility for their own education to emphasize developing ethical skills.

It isn't only organizations that can help re-scope the job of ML engineering. Educators can include social considerations in class alongside technical examples and assignments. Professional organizations can emphasize ethics outside of a formal Code: along with educators, they shape the perception of what it means to be an excellent engineer by selecting leadership, endorsing curricula, promoting best practices, and publishing or publicizing high-quality work. To support the larger project of ethical ML, professional organizations and educators must include social aspects as part of their framing of engineering success. Note that a quality framing may help here. Even highly ethically sensitive participants in these studies indicated or demonstrated a preference for talking about ethical issues as quality issues. I recommend that educators, professional organizations, and firms include high standards for ethics in their definition of high-quality engineering.

5.2 Future Work

Ethical sensitivity offers a new perspective from which to study ethical work in technology. I see particular potential to build on the work presented here in four areas: the nature of ethical sensitivity, ethical sensitivity in other technologists, ethical sensitivity in groups, and methods for studying ethical sensitivity.

5.2.1 ML Ethics Tool

I hope that practitioners, educators, and managers find the ML ethics tool described in Chapter4 to be useful. It may be help engineers find technical mitigations, understand relevant problem features, and develop particularization and judgment skills. It may give managers a free resource to train newly hired engineers in relevant mitigations (and convey the priority of ethical ML in the role and team). If the summaries are trusted, it would allow engineers could also use it to discuss their options or judgment with non-experts who don't need (and may not understand) the technical detail in academic papers or code– the credible sources engineers currently rely on to select mitigations.

Outside the firm, the ML ethics tool may help researchers get the word out about their ethical mitigations, including among people who may not have access to academic articles or conferences and less experienced or less technical players in the ML pipeline who don't have practice interpreting academic papers. It can also be used for teaching, both for teaching ML ethics specifically and for integrating ML ethics into general ML curricula. The tool can be used in curricula depending on the goal of the class and assignment. To improve familiarity with the technical mitigation options, students could be assigned to review the available options for technical ethical mitigations with a generic search task (like a scavenger hunt). To improve particularization and judgment skills, students could be assigned a problem-focused search task (e.g., find three potential mitigations for the following problem, compare them, and select the best suited). To apply their understanding of the math behind ML technology, improve their skills at reading technical academic papers, and contribute to their professional community, students could be given or asked to find an academic paper about an ethical mitigation that does not have a tool profile and submit a tool profile summarizing it.

The usefulness of this tool depends on the extent to which it is used by readers and

is updated with high quality tool profiles. The study described in this dissertation did not address these kinds of adoption or quality. I hope to present the paper at a conference (perhaps CSCW) and I will take every opportunity to offer lightning talks or CRAFT sessions like those offered at FAccT. I will work with educators to develop assignments for their students as well. Sometime after the conference presentation, at least one successful cohort of student assignments, and some further development on the search tool (as described in Chapter 4) I will consider developing some firm-facing messaging and support resource (like a training guide.) I hope that these efforts will encourage researchers and students to contribute high-quality tool profiles and spread awareness to practitioners.

5.2.2 Ethical Sensitivity

This dissertation offers some rich descriptive data about ethical sensitivity in facial recognition training data curation, but doesn't offer generalizable findings about other circumstances or how to encourage it across work settings. This is appropriate for ethical sensitivity's stage of development. There is a lot left to learn about ethical sensitivity in technology development and in general. Although there's a lot of literature about ethical sensitivity, it varies quite widely in how it's been conceptualized and operationalized [24, 106, 184, 186]. Review authors note disagreements about "its definition, its characteristics, the conditions needed for it to occur, or the outcomes to professionals and society" [184] and researchers cite these conflicts as sources of inconsistent and confusing study results [137]. This may come from a lack of preliminary qualitative work describing ethical sensitivity's features in favor of immediately developing scales to measure it in structured interviews [19] and surveys [143]. Although surveys are appealing for their ability to scale to entire organizations or a substantial number of members of a professional organization, I do not believe it's appropriate to use them to measure ethical sensitivity given the lack of

consensus on how to measure it, or even what it is. I recommend that more rich data about ethical sensitivity be collected in different circumstances, including a variety of ethical issues, occupations, organizations, and countries, before attempting to generalize about it. This data can conflict with or add nuance to the three-activity model presented in Chapter 2, improving our understanding and ability to eventually generalize.

This study understood ethical sensitivity as a skill, not as an immutable characteristic of a person. Future work can test this assumption using research about skill development and education, giving us a better picture of ES and revealing how to develop it in students, trainees, and professionals. One example of research that may help us understand and improve ethical sensitivity as a skill is deliberate practice.

We know from studies in other fields that deliberate practice can improve many skills, including many cognitive ones. Deliberate practice requires performing a skill repeatedly with focused concentration and with direct, specific feedback. This often means breaking a large skill (e.g., basketball) down to component parts (e.g., free throws, dribbling, passing)– simple experience (e.g., playing many games of basketball) can't get you as far as deliberate practice can [59]. So can we break ES down into component parts, offer feedback, and encourage focused concentration during repetition? Does deliberate practice help? This study offers a place to start when selecting component parts in Chapter 2: recognition, particularization, and judgment. Particularization is especially complex, and may need to be broken down further.

Particularization

Prior work on particularization lacks agreement on its features and methods for observing it, and in my view underestimates its importance. Technology development as a setting, think aloud as a method, and intellectual traditions that center practices, power, and context

offer a unique opportunity to understand particularization and its connection to judgment.

Immediately after recognizing an ethical issues, participants seemed to prefer reflection over information seeking (in Chapter 3) but they did demonstrate they were adept at information seeking, especially once they were prompted to do so the study described in Chapter 4. Information seeking is an area of particular interest in Information Science and also offers an important opportunity intervene to inform engineers about values, interventions, and social implications of technical choices while they are building understanding. Further work can shed more light on information seeking during particularization, especially what task, contextual, and personal factors can cause or encourage information seeking.

Participants frequently sought and selected high-level resources from Medium, Towards Data Science, and similar sources, using summaries to orient themselves to the larger landscape of research and to find academic research articles. The reliance on conference and journal articles was surprising to me and may be encouraging to researchers. Participants' use of these articles can offer some guidance to their writers as well: to the extent that it's possible, highlight the specific purpose, requirements, how it works, and as many elements of fit as are relevant in the abstract. Continue to write summary blog posts and include working examples of code when you can. If there are circumstances where your intervention will not work or could result in unintended consequences, make that information as clear as you can by putting it alongside the types of information engineers are looking for: don't assume they will read the entire article or blog post from beginning to end.

5.2.3 Ethical sensitivity in broader technology development

ML engineers are of course not the only people involved in technology development whose ethical sensitivity is interesting. As I discussed in the introduction to this chapter, regulators, managers, governments, users, institutions and many others also contribute to

the proliferation of ML-driven products and the harms that these technologies can create or propagate. To understand and prevent harm we need to understand what causes these people and groups to recognize, particularize, and judge as well. This will not only improve our understanding and ability to intervene throughout the ML ecosystem, but also allow us compare other actors' ethical sensitivity to that of ML engineers. Perhaps there are circumstances, norms, values, incentives or other factors that differ among professions or people that can help us understand how ES is developed and distributed.

Second, there are people in positions analogous to ML engineers, but who work on other types of ethically important technology. Smart devices and mobile phones have intimate data about people's lives, bodies, living spaces, and more: do the people who make ostensibly technical decisions about when and how that data is collected, stored, and used recognize the ethical implications of their work? When someone at a company with sensitive data (say, a credit reporting company) becomes aware of a security flaw, what do they do next? When the press questions their company about radicalization, privacy, or fairness, how do people decide where to go next? What do technology developers have in common with one another and how do they differ across and within disciplines.

5.2.4 Ethical Sensitivity in groups

When studying ethical behavior in organizations, individuals can only take you so far. In order to answer important questions like the last one in the list above, you need to understand how ethical sensitivity operates in groups.

Do existing hierarchical and political structures dictate how groups respond when a person recognizes an ethical issue? Do some organizational structures encourage more ethical recognition, robust particularization, or effective judgments? Does the context of the work– for profit, governmental, academic, or free/open-source– alter recognition, particu-

larization, or judgment? Once a group agrees that there's a potential ethical issue, are there patterns in the kinds of reflection, discussion, and information seeking that happen? Given some goal (like quantity of discussion, effectiveness of eventual judgment, or an ethical standard) what kinds of particularization lead to success?

Theories of group interaction may help guide this research agenda. We can test, for example, whether "groupthink" interferes with particularization and what techniques work to disrupt that [103]. Does it matter whether recognition occurs early in a groups formation, during a period of conflict, as they are disbanding [177]?

The effect of management behavior, job descriptions, and evaluation schema on recognition is of particular interest to me.

5.2.5 Methods for studying ethical sensitivity

To answer these questions, we need to build out a robust suite of tools for observing ethical sensitivity in technology development. I have some recommendations for methods development.

1. Tools should be as situated in a real work environment as they can be. The method presented here, although an improvement in this regard over prior methods, does not capture any organizational context, which I expect will be interesting and explanatory.

2. Novel methods should be developed to observe ethical sensitivity in groups small large groups.

3. Methods that support different scales of study should be developed in order to allow for more robust comparison. Once we better understand ethical sensitivity in technology development and group influences, quantitative methods will help us answer comparative and causal questions.

This dissertation contributed a review of interdisciplinary research about ethical sensitivity, argued for its study in technology development, developed and employed a method for observing individual ethical sensitivity in ML engineers working with unfamiliar training data, and designed a tool that may help machine learning engineers who have recognized an ethical problem to particularize and judge more effectively. It suggests that ML engineers are largely ethically sensitive, and that relatively minor adjustments (like asking them whether there are potential ethical problems in their work and clarifying that ethical concerns are within their job description) could go a long way toward catching ML harm upstream.

Appendix A: Ethical Sensitivity Corpus

1. Ahn, S.H., & Yeom, H.A. (2014). Moral sensitivity and critical thinking disposition of nursing students in Korea. International Journal of Nursing Practice, 20(5), 482–489. https://doi.org/ 10.1111/ijn.12185

2. Akaah, I. P. (1989). Differences in research ethics judgments between male and female marketing professionals. Journal of Business Ethics, 8(5), 375–381. https://doi.org/ 10.1007/BF00381729

3. Akabayashi, A., Slingsby, B. T., Kai, I., Nishimura, T., & Yamagishi, A. (2004). The development of a brief and objective method for evaluating moral sensitivity and reasoning in medical students. BMC Medical Ethics, 5(1), 1. https://doi.org/10.1186/1472-6939-5-1

4. Al-Kazemi, A. A., & Zajac, G. (1999). Ethics Sensitivity and Awareness Within Organizations in Kuwait: An Empirical Exploration of Espoused Theory and Theory-in-Use. Journal of Business Ethics, 9.

5. Ameen, E. C., Guffey, D. M., & McMillan, J. J. (1996). Gender differences in determining the ethical sensitivity of future accounting professionals. Journal of Business Ethics, 15(5), 591–597. https://doi.org/10.1007/BF00381934

6. Amiri, E., Ebrahimi, H., Vahidi, M., Asghari Jafarabadi, M., & Namdar Areshtanab, H. (2019). Relationship between nurses' moral sensitivity and the quality of care. Nursing Ethics, 26(4), 1265–1273. https://doi.org/10.1177/0969733017745726

7. Barnett, T., & Valentine, S. (2004). Issue contingencies and marketers' recognition of ethical issues, ethical judgments and behavioral intentions. Journal of Business

Research, 57(4), 338–346. https://doi.org/10.1016/S0148-2963(02)00365-X

8. Basar, Z., & Cilingir, D. (2019). Evaluating ethical sensitivity in surgical intensive care nurses. Nursing Ethics, 26(7–8), 2384–2397. https://doi.org/10.1177/0969733018792739

9. Baykara, Z. G., Demir, S. G., & Yaman, S. (2015). The effect of ethics training on students recognizing ethical violations and developing moral sensitivity. Nursing Ethics, 22(6), 661–675. https://doi.org/10.1177/0969733014542673

10. Bebeau, M. J., & Brabeck, M. M. (1987). Integrating Care and Justice Issues in Professional Moral Education: A Gender Perspective. Journal of Moral Education, 16(3), 189–203. https://doi.org/10.1080/0305724870160304

11. Bebeau, M., Rest, J., & Yamoor, C. (1985). Measuring dental students' ethical sensitivity. Journal of Dental Education, 49, 225–235.

12. Blodgett, J. G., Lu, L.-C., Rose, G. M., & Vitell, S. J. (2001). Ethical Sensitivity to Stakeholder Interests: A Cross-Cultural Comparison. Journal of the Academy of Marketing Science, 29(2), 190–202. https://doi.org/10.1177/03079459994551

13. Borhani, F., Abbaszadeh, A., Mohamadi, E., Ghasemi, E., & Hoseinabad-Farahani, M. J. (2017). Moral sensitivity and moral distress in Iranian critical care nurses. Nursing Ethics, 24(4), 474–482. https://doi.org/10.1177/0969733015604700

14. Butterfield, K. D., Trevin, L. K., & Weaver, G. R. (2000). Moral Awareness in Business Organizations: Influences of Issue-Related and Social Context Factors. Human Relations, 53(7), 981–1018. https://doi.org/10.1177/0018726700537004

15. Chia, A., & Mee, L. S. (2000). The Effects of Issue Characteristics on the Recognition of Moral Issues. Journal of Business Ethics, 16.

16. Clarkeburn, H. (2002). A Test for Ethical Sensitivity in Science. Journal of Moral Education, 31(4), 439–453. https://doi.org/10.1080/0305724022000029662

17. Claypool, G. A., Fetyko, D. F., & Pearson, M. A. (1990). Reactions to ethical

dilemmas: A study pertaining to certified public accountants. Journal of Business Ethics, 9(9), 699–706. https://doi.org/10.1007/BF00386352

18. Comrie, R. W. (2012). An analysis of undergraduate and graduate student nurses' moral sensitivity. Nursing Ethics, 19(1), 116–127. https://doi.org/ 10.1177/0969733011411399

19. Dalla Nora, C. R., Zoboli, E. L., & Vieira, M. M. (2019). Validation of a Brazilian version of the moral sensitivity questionnaire. Nursing Ethics, 26(3), 823–832. https://doi.org/ 10.1177/0969733017720849

20. Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The Contribution of Emotion and Cognition to Moral Sensitivity: A Neurodevelopmental Study. Cerebral Cortex, 22(1), 209–220. https://doi.org/10.1093/cercor/bhr111

21. Dotger, B. H. (2010). "I Had No Idea": Developing Dispositional Awareness and Sensitivity through a Cross-Professional Pedagogy. Teaching and Teacher Education: An International Journal of Research and Studies, 26(4), 805–812. https://doi.org /10.1016/j.tate.2009.10.017

22. Ersoy, N., & Göz, F. (2001). The Ethical Sensitivity of Nurses in Turkey. Nursing Ethics, 8(4), 299–312. https://doi.org/ 10.1177/096973300100800403

23. Ersoy, N., & Gündoğmuş, Ü. N. (2003). A Study of the Ethical Sensitivity of Physicians in Turkey. Nursing Ethics, 10(5), 472–484. https://doi.org/ 10.1191/0969733003ne6290a

24. Erwin, W. J. (2000). Supervisor Moral Sensitivity. Counselor Education and Supervision, 40(2), 115–127. https://doi.org/10.1002/j.1556-6978.2000.tb01243.x

25. Escolar-Chua, R. L. (2018). Moral sensitivity, moral distress, and moral courage among baccalaureate Filipino nursing students. Nursing Ethics, 25(4), 458–469. https://doi.org/ 10.1177/0969733016654317

26. Fiolleau, K., & Kaplan, S. E. (2017). Recognizing Ethical Issues: An Examination of Practicing Industry Accountants and Accounting Students. Journal of Business Ethics, 142(2), 259–276. https://doi.org/10.1007/s10551-016-3154-2

27. Fowler, S., Zeidler, D., & Sadler, T. (2009). Moral Sensitivity in the Context of Socioscientific Issues in High School Science Students. International Journal of Science Education, 31(2), 279–296.

28. Gholami, K., Kuusisto, E., & Tirri, K. (2015). Is ethical sensitivity in teaching culturally bound? Comparing Finnish and Iranian teachers' ethical sensitivity. Compare: A Journal of Comparative and International Education, 45(6), 886–907. https://doi.org/10.1080/03057925.2014.984588

29. González-de Paz, L., Kostov, B., Sisó-Almirall, A., & Zabalegui-Yárnoz, A. (2012). A Rasch analysis of nurses' ethical sensitivity to the norms of the code of conduct. Journal of Clinical Nursing, 21(19pt20), 2747–2760. https://doi.org/10.1111/j.1365-2702.2012.04137.x

30. Gwak, S., Kim, H., & Lee, W. (2013). Development and Application of the Information Ethical Sensitivity Measurement Tool for College Students in Korea. 2013 IEEE 37th Annual Computer Software and Applications Conference Workshops, 639–643. https://doi.org/10.1109/COMPSACW.2013.104

31. Han, S. S., Kim, J., Kim, Y. S., & Ahn, S. (2010). Validation of a Korean version of the Moral Sensitivity Questionnaire. Nursing Ethics, 17(1), 99–105. https://doi.org/10.1177/0969733009349993

32. Harenski, C. L., Antonenko, O., Shane, M. S., & Kiehl, K. A. (2008). Gender differences in neural mechanisms underlying moral sensitivity. Social Cognitive and Affective Neuroscience, 3(4), 313–321. https://doi.org/10.1093/scan/nsn026

33. Hebert, P. C., Meslin, E. M., & Dunn, E. V. (1992). Measuring the ethical sensitivity of medical students: A study at the University of Toronto. Journal of Medical Ethics, 18(3), 142–147. https://doi.org/10.1136/jme.18.3.142

34. Hebert, P., Meslin, E. M., Dunn, E. V., Byrne, N., & Reid, S. R. (1990). Evaluating ethical sensitivity in medical students: Using vignettes as an instrument. Journal of Medical

Ethics, 16(3), 141–145. https://doi.org/10.1136/jme.16.3.141

35. Heggestad, A. K. T., Nortvedt, P., & Slettebøo, Å. (2013). The importance of moral sensitivity when including persons with dementia in qualitative research. Nursing Ethics, 20(1), 30–40. https://doi.org/10.1177/0969733012455564

36. Hemberg, J., & Bergdahl, E. (2020). Ethical sensitivity and perceptiveness in palliative home care through co-creation. Nursing Ethics, 27(2), 446–460. https://doi.org/10.1177/0969733019849464

37. Hollingworth, D., & Valentine, S. (2015). The Moderating Effect of Perceived Organizational Ethical Context on Employees' Ethical Issue Recognition and Ethical Judgments. Journal of Business Ethics, 128(2), 457–466. https://doi.org/10.1007/s10551-014-2088-9

38. Huang, F. F., Yang, Q., Zhang, J., Khoshnood, K., & Zhang, J. P. (2016). Chinese nurses' perceived barriers and facilitators of ethical sensitivity. Nursing Ethics, 23(5), 507–522. https://doi.org/10.1177/0969733015574925

39. Huang, F. F., Yang, Q., Zhang, J., Zhang, Q. H., Khoshnood, K., & Zhang, J. P. (2016). Cross-cultural validation of the moral sensitivity questionnaire-revised Chinese version. Nursing Ethics, 23(7), 784–793. https://doi.org/10.1177/0969733015583183

40. Huang, H., Ding, Y., Wang, H., Khoshnood, K., & Yang, M. (2018). The Ethical Sensitivity of Health Care Professionals Who Care For Patients Living With HIV Infection in Hunan, China: A Qualitative Study. Journal of the Association of Nurses in AIDS Care, 29(2), 266–274. https://doi.org/10.1016/j.jana.2017.09.001

41. Ineichen, C., Christen, M., & Tanner, C. (2017). Measuring value sensitivity in medicine. BMC Medical Ethics, 18(1), 1–12. https://doi.org/10.1186/s12910-016-0164-7

42. Jalili, F., Saeidnejad, Z., & Aghajani, M. (2020). Effects of spirituality training on the moral sensitivity of nursing students: A clinical randomized controlled trial: Clinical Ethics. https://doi.org/10.1177/1477750919898346

43. Karcher, J. N. (1996). Auditors' ability to discern the presence of ethical problems. Journal of Business Ethics, 15(10), 1033–1050. https://doi.org/10.1007/BF00412045

44. Khalighi, E., Solaimanizadeh, L., Borji, M., Tarjoman, A., Soltany, B., & Zareie, F. (2020). Investigating relationship between religious commitment and moral sensitivity in nurses working in ICU. BMC Research Notes, 13(1), 41. https://doi.org/10.1186/s13104-020-4912-x

45. Kidwell, J. M., Stevens, R. E., & Bethke, A. L. (1987). Differences in ethical perceptions between male and female managers: Myth or reality? Journal of Business Ethics, 6(6), 489–493. https://doi.org/10.1007/BF00383291

46. Kim, W. J., & Park, J. H. (2019). The effects of debate-based ethics education on the moral sensitivity and judgment of nursing students: A quasi-experimental study. Nurse Education Today, 83, 104200. https://doi.org/10.1016/j.nedt.2019.08.018

47. Kim, Y. S., Kang, S. W., & Ahn, J. A. (2013). Moral sensitivity relating to the application of the code of ethics. Nursing Ethics, 20(4), 470–478. https://doi.org/10.1177/0969733012455563

48. Kim, Y. S., Park, J. W., You, M. A., Seo, Y. S., & Han, S. S. (2005). Sensitivity to ethical issues confronted by Korean hospital staff nurses. Nursing Ethics, 12(6), 595–605. https://doi.org/10.1191/0969733005ne829oa

49. Kirilmaz, H., Akbolat, M., & Kahraman, G. (2015). A Research about the Ethical Sensitivity of Healthcare Professionals. International Journal of Health Sciences (IJHS), 3(3). https://doi.org/10.15640/ijhs.v3n3a7

50. Kulju, K., Suhonen, R., & Leino-Kilpi, H. (2013). Ethical problems and moral sensitivity in physiotherapy: A descriptive study. Nursing Ethics, 20(5), 568–577. https://doi.org/10.1177/0969733012468462

51. Lee, E., & Kim, Y. (2020). The relationship of moral sensitivity and patient safety attitudes with nursing students' perceptions of disclosure of patient safety inci-

dents: A cross-sectional study. PLOS ONE, 15(1), e0227585. https://doi.org/ 10.1371/journal.pone.0227585

52. Lee, H. L., Huang, S.-H., & Huang, C.-M. (2017). Evaluating the effect of three teaching strategies on student nurses' moral sensitivity. Nursing Ethics, 24(6), 732–743. https://doi.org/10.1177/0969733015623095

53. Lind, R. (1997). Ethical Sensitivity in Viewer Evaluations of a TV News Investigative Report. Human Communication Research, 23(4), 535–561. https://doi.org/10.1111/j.1468-2958.1997.tb00409.x

54. Lind, R. A., & Rarick, D. L. (1999). Viewer Sensitivity to Ethical Issues in TV Coverage of the Clinton-Flowers Scandal. Political Communication, 16(2), 169–181. https://doi.org/10.1080/105846099198712

55. Lind, R. A., & Swenson-Lepper, T. (2013). Measuring Sensitivity to Conflicts of Interest: A Preliminary Test of Method. Science and Engineering Ethics, 19(1), 43–62. https://doi.org/10.1007/s11948-011-9319-6

56. Lind, R. A., Swenson-Lepper, T., & Rarick, D. L. (1998). Identifying patterns of ethical sensitivity in TV news viewers: An assessment of some critical viewing skills. Journal of Broadcasting & Electronic Media, 42(4), 507–519. https://doi.org/10.1080/08838159809364465

57. Liu, J., Yuan, B., Luo, Y., & Cui, F. (2019). Intrinsic functional connectivity of medial prefrontal cortex predicts the individual moral bias in economic valuation partially through the moral sensitivity trait. Brain Imaging and Behavior. https://doi.org/10.1007/s11682-019-00152-1

58. Lützén, K., Blom, T., Ewalds-Kvist, B., & Winch, S. (2010). Moral stress, moral climate and moral sensitivity among psychiatric professionals. Nursing Ethics, 17(2), 213–224. https://doi.org/10.1177/0969733009351951

59. Lützén, K., Dahlqvist, V., Eriksson, S., & Norberg, A. (2006). Developing the

Concept of Moral Sensitivity in Health Care Practice. Nursing Ethics, 13(2), 187–196. https://doi.org/ 10.1191/0969733006ne837oa

60. Lützén, K., Evertzon, M., & Nordin, C. (1997). Moral Sensitivity in Psychiatric Practice. Nursing Ethics, 4(6), 472–482. https://doi.org/10.1177/096973309700400604

61. Lützén, K., Johansson, A., & Nordström, G. (2000). Moral Sensitivity: Some differences between nurses and physicians. Nursing Ethics, 7(6), 520–530. https://doi.org/ 10.1177/096973300000700607

62. Martinov-Bennie, N., & Mladenovic, R. (2015). Investigation of the Impact of an Ethical Framework and an Integrated Ethics Education on Accounting Students' Ethical Sensitivity and Judgment. Journal of Business Ethics, 127(1), 189–203. https://doi.org/ 10.1007/s10551-013-2007-5

63. Mert Boğa, S., Aydin Sayilan, A., Kersu, Ö., & Baydemİr, C. (2020). Perception of care quality and ethical sensitivity in surgical nurses. Nursing Ethics, 27(3), 673–685. https://doi.org/10.1177/0969733020901830

64. Milliken, A., Ludlow, L., DeSanto-Madeya, S., & Grace, P. (2018). The development and psychometric validation of the Ethical Awareness Scale. Journal of Advanced Nursing, 74(8), 2005–2016. https://doi.org/10.1111/jan.13688

65. Milliken, A., Ludlow, L., & Grace, P. (2019). Ethical Awareness Scale: Replication Testing, Invariance Analysis, and Implications. AJOB Empirical Bioethics, 10(4), 231–240. https://doi.org/10.1080/23294515.2019.1666176

66. Molenberghs, P., Gapp, J., Wang, B., Louis, W. R., & Decety, J. (2016). Increased Moral Sensitivity for Outgroup Perpetrators Harming Ingroup Members. Cerebral Cortex, 26(1), 225–233. https://doi.org/10.1093/cercor/bhu195

67. Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourão-Miranda, J., Andreiuolo, P. A., & Pessoa, L. (2002). The Neural Correlates of Moral Sensitivity: A Functional Magnetic Resonance Imaging Investigation of Basic and Moral Emotions. The

Journal of Neuroscience, 22(7), 2730–2736. https://doi.org/10.1523/JNEUROSCI.22-07-02730.2002

68. Moll, J., Oliveira-Souza, R. de, Garrido, G. J., Bramati, I. E., Caparelli-Daquer, E. M. A., Paiva, M. L. M. F., Zahn, R., & Grafman, J. (2007). The self as a moral agent: Linking the neural bases of social agency and moral sensitivity. Social Neuroscience, 2(3–4), 336–352. https://doi.org/10.1080/17470910701392024

69. Morton, K. R., Worthley, J. S., Testerman, J. K., & Mahoney, M. L. (2006). Defining features of moral sensitivity and moral motivation: Pathways to moral reasoning in medical students. Journal of Moral Education, 35(3), 387–406. https://doi.org/10.1080/03057240600874653

70. Muramatsu, T., Nakamura, M., Okada, E., Katayama, H., & Ojima, T. (2019). The development and validation of the Ethical Sensitivity Questionnaire for Nursing Students. BMC Medical Education, 19(1), 215. https://doi.org/10.1186/s12909-019-1625-8

71. Myyry, L., & Helkama, K. (2002). The Role of Value Priorities and Professional Ethics Training in Moral Sensitivity. Journal of Moral Education, 31(1), 35–50. https://doi.org/ 10.1080/03057240120111427 72. Narvaez, D. F. (1996). Moral Perception: A New Construct? https://eric.ed.gov/?id=ED398636

73. Ohnishi, K., Kitaoka, K., Nakahara, J., Välimäki, M., Kontio, R., Anttila, M. (2019). Impact of moral sensitivity on moral distress among psychiatric nurses. Nursing Ethics, 26(5), 1473–1483. https://doi.org/10.1177/0969733017751264

74. Palazoğlu, C. A., & Koç, Z. (2019). Ethical sensitivity, burnout, and job satisfaction in emergency nurses. Nursing Ethics, 26(3), 809–822. https://doi.org/10.1177/0969733017720846

75. Park, M., Kjervik, D., Crandell, J., & Oermann, M. H. (2012). The relationship of ethics education to moral sensitivity and moral reasoning skills of nursing students. Nursing Ethics, 19(4), 568–580. https://doi.org/10.1177/0969733011433922

76. Patterson, D. M. (2001). Causal Effects of Regulatory, Organizational and Personal Factors on Ethical Sensitivity. Journal of Business Ethics, 37.

77. Radtke, R. R. (2000). The Effects of Gender and Setting on Accountants' Ethically Sensitive Decisions. Journal of Business Ethics.

78. Reidenbach, R. E., & Robin, D. P. (1988). Some initial steps toward improving the measurement of ethical evaluations of marketing activities. Journal of Business Ethics, 871–879.

79. Reidenbach, R. E., & Robin, D. P. (1990). Toward the Development of a Multidimensional Scale for Improving Evaluations of Business Ethics. Journal of Business Ethics, 15.

80. Reidenbach, R. E., Robin, D. R., & Dawson, L. (1991). An application and extension of a multidimensional ethics scale to selected marketing practices and marketing groups. Journal of the Academy of Marketing Science, 10.

81. Reynolds, S. J. (2006). Moral awareness and ethical predispositions: Investigating the role of individual differences in the recognition of moral issues. Journal of Applied Psychology, 91(1), 233–243. https://doi.org/10.1037/0021-9010.91.1.233

82. Robertson, D., Snarey, J., Ousley, O., Harenski, K., Bowman, F. D., Gilkey, R., & Kilts, C. (2007). The neural processing of moral sensitivity to issues of justice and care. Neuropsychologia, 45(4), 755–766. https://doi.org/10.1016/j.neuropsychologia.2006.08.014

83. Roshanzadeh, M., Vanaki, Z., & Sadooghiasl, A. (2019). Sensitivity in ethical decision-making: The experiences of nurse managers. Nursing Ethics, 0969733019864146. https://doi.org/ 10.1177/0969733019864146

84. Sadler, T. D. (2004). Moral sensitivity and its contribution to the resolution of socio-scientific issues. Journal of Moral Education, 33(3), 339–358. https://doi.org/10.1080/0305724042000733

85. Sahin, S. Y., Iyigun, E., & Acikel, C. (2015). Validity and Reliability of a Turkish Version of the Modified Moral Sensitivity Questionnaire for Student Nurses. Ethics

Behavior, 25(4), 351–359. https://doi.org/10.1080/10508422.2014.948955

86. Savage, J. S., & Favret, J. O. (2006). Nursing students' perceptions of ethical behavior in undergraduate nursing faculty. Nurse Education in Practice, 6(1), 47–54. https://doi.org/10.1016/j.nepr.2005.08.002

87. Schutte, I., Wolfensberger, M., & Tirri, K. (2014). The Relationship between Ethical Sensitivity, High Ability and Gender in Higher Education Students. Gifted and Talented International, 29(1–2), 39–48. https://doi.org/10.1080/15332276.2014.11678428

88. Sedgwick, M., Yanicki, S., & Pijl, E. M. (2020). Analysis of Undergraduate Nursing Students39; Sensitivity to Microethical Dilemmas During Simulation. Journal of Nursing Education, 59(2), 88–92. https://doi.org/10.3928/01484834-20200122-06

89. Shaub, M. K., Finn, D. W., & Munter, P. (1993). The effects of auditors' ethical orientation on commitment and ethical sensitivity. Behavioral Research in Accounting, 5(1), 145–169.

90. Shawver, T. J., & Sennetti, J. T. (2009). Measuring Ethical Sensitivity and Evaluation. Journal of Business Ethics, 88(4), 663–678. https://doi.org/10.1007/s10551-008-9973-z

91. Sidani, Y., Zbib, I., Rawwas, M., & Moussawer, T. (2009). Gender, age, and ethical sensitivity: The case of Lebanese workers. Gender in Management; Bradford, 24(3), 211–227. http://dx.doi.org/10.1108/17542410910950886

92. Simga-Mugan, C., Daly, B. A., Onkal, D., & Kavut, L. (2005). The Influence of Nationality and Gender on Ethical Sensitivity: An Application of the Issue-Contingent Model. Journal of Business Ethics, 57(2), 139–159. https://doi.org/10.1007/s10551-004-4601-z

93. Sirin, S. R., Brabeck, M. M., Satiani, A., & Rogers-Serin, L. (2003). Validation of a Measure of Ethical Sensitivity and Examination of the Effects of Previous Multicultural and Ethics Courses on Ethical Sensitivity. Ethics & Behavior, 13(3), 221–235.

94. Sparks, J. R. (2015). A social cognitive explanation of situational and individual effects on moral sensitivity. Journal of Applied Social Psychology, 45(1), 45–54. https://doi.org/10.1111/jasp.12274

95. Sparks, J. R., & Hunt, S. D. (1998). Marketing Researcher Ethical Sensitivity: Conceptualization, Measurement, and Exploratory Investigation. Journal of Marketing, 62(2), 92–109. https://doi.org/10.1177/002224299806200207

96. Swenson-Lepper, T. (2005). Ethical Sensitivity for Organizational Communication Issues: Examining Individual and Organizational Differences. Journal of Business Ethics, 59(3), 205–231. https://doi.org/10.1007/s10551-005-2925-y

97. Szabó, C., Németh, A., & Kéri, S. (2013). Ethical sensitivity in obsessive-compulsive disorder and generalized anxiety disorder: The role of reversal learning. Journal of Behavior Therapy and Experimental Psychiatry, 44(4), 404–410. https://doi.org/10.1016/j.jbtep.2013.04.001

98. Tirri, K., & Nokelainen, P. (2007). Comparison of Academically Average and Gifted Students' Self-Rated Ethical Sensitivity. Educational Research and Evaluation, 13(6), 587–601. https://doi.org/10.1080/13803610701786053

99. Tirri, K., Nokelainen, P., & Holm, K. (2008). Ethical Sensitivity of Finnish Lutheran 7th- 9th Grade Students. Getting Involved.

100. Tuvesson, H., & Lützén, K. (2017). Demographic factors associated with moral sensitivity among nursing students. Nursing Ethics, 24(7), 847–855. https://doi.org/ 10.1177/0969733015

101. Walsh, L., Onorato, M., & Simms, S. (2016). Ethical Sensitivity and Its Relationship to Personality and Area of Study. SAM Advanced Management Journal, 81(2), 11–20.

102. Wittmer, D. (1992). Ethical Sensitivity and Managerial Decisionmaking: An Experiment. Journal of Public Administration Research and Theory, 2(4), 443–462. https://doi.org/10.1093/oxfordjournals.jpart.a037147

103. Wittmer, D. P. (2000). Ethical Sensitivity in Management Decisions: Developing and Testing a Perceptual Measure Among Management and Professional Student Groups. Teaching Business Ethics, 4(2), 181–205. https://doi.org/10.1023/A:1009866315139

104. Xiang, Y., Cao, Y., & Dong, X. (2020). Childhood maltreatment and moral sensitivity: An interpretation based on schema theory. Personality and Individual Differences, 160, 109924. https://doi.org/10.1016/j.paid.2020.109924

105. Yeom, H. A., Ahn, S. H., & Kim, S. J. (2017). Effects of ethics education on moral sensitivity of nursing students. Nursing Ethics, 24(6), 644–652. https://doi.org/10.1177/0969733015622060

106. Yetmar, S. A., & Eastman, K. K. (2000). Tax Practitioners' Ethical Sensitivity: A Model and Empirical Examination. Journal of Business Ethics, 18.

107. Zhang, N., Li, J., Xu, Z., & Gong, Z. (2019). A latent profile analysis of nurses' moral sensitivity. Nursing Ethics, 0969733019876298. https://doi.org/10.1177/0969733019876298

108. Zhang, N., & Zhang, J. (2016). Business ethical sensitivity of Chinese insurance agents: Scale development and validation. International Journal of Information Systems and Change Management, 8, 3. https://doi.org/10.1504/IJISCM.2016.077943

Appendix B: Script

Hi, good to [see/meet] you! I'm Karen. I'm from the University of Maryland, College Park's School of Information and we are interested in how you think about and work with training data.

Thanks for participating in this study. My goal here is to learn more about how ML engineers work with unfamiliar training data so that we can make the process of exploring, manipulating, augmenting, and labelling to serve your purposes better. I'll be reading from a script for a lot of our time today, just to make sure participants have a similar experience, but you can feel free to ask questions whenever.

Is it OK with you if I start recording?

First, I'll ask you a couple questions about your work. In what capacity do you work with machine learning algorithms and training data?

Did you take any formal courses on ML?

How many years of experience do you have with ML?

Thanks!

This study consists of two tasks. The first one will take 25 minutes. I'm going to send you a link. It has an ML problem and a sample set of training data. Send link.

Alright, I'll ask you to think out loud as you review the data and come up with a plan for whether and how to use it for addressing the ML problem. You are welcome to download the data and play with it on your machine, look at it online, whatever you would do if you

were considering a training data set and a new ML problem. Say aloud whatever comes to your mind– what you are paying attention to and thinking about– rather than explaining how to do what you are doing or what you think you should do. The goal here is to get a sense of how you naturally work with unfamiliar training data, so everything you have access to is fair game: the internet, a book, contents of your computer. If I'm confused about anything, I'll ask you about it when we are done, so don't worry about being clear. If you're quiet for a while, I might jump in and remind you to keep thinking aloud, because what you are referring to and thinking about are important information for us to be able to improve your tools. I'm not as interested in the answer to the problem as I am about how you are thinking about while working.

I didn't want to overwhelm the Drive or your machine with tens of thousands of images, so if there's anything you would like to do, but can't do with this sample, you can include that in your plan (assuming you'll get the entire dataset sometime in the future).

Do you have any questions?

Alright, go ahead and get started. Start timer for 25 minutes.

If participant has been silent more than 20 seconds, prompt with one of the following:

"What are you thinking?"

"Don't forget to think aloud as you work"

"What's going through your mind as you do this?"

"What's standing out to you here?"

Alright, 25 minutes is up! Thanks for your help. Let's talk a little bit about what you'd do next.

First, can you describe what your approach is?

What would your next steps be?

How would you approach labelling?

What would an ML model trained on this model be useful for?

What would it not be useful for?

Would you want any other kinds of data to improve the model?

Did you notice any potential ethical or legal issues in the problem or data?

Do you have experience with similar data or ML problems, or was this mostly a new domain?

Great!

[If they did not notice the potential bias problem] So for the next step, I'll ask you to imagine that after a few weeks of working with this data, you and your team noticed that there were a lot more men than women and that there were some skin tones missing. You're worried that this will hurt the performance of the model for those groups.

[If they did mention the potential bias problem] So, it sounds like you already noticed this, but there were a lot more men than women and that there were some skin tones missing. You're worried that this will hurt the performance of the model for those groups.

Go ahead and continue to think aloud as you use the internet, your own resources, or reflection to decide how you might move forward knowing this.

I wanted to apologize for not mentioning that we were interested in ethics for this study when we started. In addition to how people work with unfamiliar training data, I also wanted to see if anyone noticed the issue and, if they did, how they reacted, but I couldn't do that if everyone knew we were looking at ethics from the beginning. I certainly don't think that not noticing something in the first 30 minutes means that you never would have, nor am I convinced that ethics gatekeeper is part of your job, but we are hoping to make that noticing easier by testing a tool that might. Half of participants got this tool that we are trying to improve. Does this all make sense? Your name won't be associated with the data you provided at all, but knowing that I wasn't upfront with you, you have the right to

withdraw your data without losing your incentive. Would you like to do that?

OK, thanks for understanding.

[IF they say that they wouldn't build this thing at all] So i hear you saying that you would choose not to build this at all– that makes a lot of sense.

First, let's talk a little about that. What signaled to you that this might not be an ethical project?

Other Follow-up Questions for use if there's time, or if the participant declined to build

Have you ever noticed an ethical issue during your work? What did you do?

Have you ever been unsure about whether to go ahead, or whether to implement some kind of ethical intervention? What caused you to notice that issue, and what did you do?

Where would you go for information if you weren't sure about the ethics of something, or to decide what to do?

What interventions for ethical issues (of any kind) are you aware of?

What pieces of information about an ethical issue are important when you are looking for a solution?

What kind of information do you look up or ask someone about?

What sources for information about ethical issues and interventions do you trust?

Appendix C: Datasheet

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created to provide images that can be used to study face detection in an unconstrained setting where image characteristics (such as pose, illumination, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The initial version of the dataset was created by researchers at AVID corporation.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number. The construction of the original database was funded by AVID corporation.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and

ratings; people and interactions between them; nodes and edges)? Please provide a description.

The dataset consists of just over 65,000 high-quality PNG images at 1024x1024 resolution. Each instance includes at least one human face. Images were crawled from Photobucket to increase the likelihood that it has good coverage of accessories, including glasses, sunglasses, make-up, hair accessories, hats, etc.

How many instances are there in total (of each type, if appropriate)?

Images: 65,104

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable)

A full-resolution sample of the data is available for download. The sample is randomly selected, and so expected to be representative of the larger dataset in terms of image characteristics.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The data consists of unprocessed images of faces.

Is there a label or target associated with each instance? If so, please provide a description.

There is no label associated with each instance.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Instances are not missing information, but metadata was stripped from the original images to preserve the privacy of Photobucket users. They do not contain labels of any kind.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

There are no links.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Data provided is not split.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There are some images with more than one face in them. All images contain at least one face, centered on the image's center pixel.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any re-

strictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

This data comprises communication that was intended to be public, but publishers (individual Photobucket users) may not have anticipated that it would be used in this way. Further, publishers may have made images available of people other than themselves without permission.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Images were not thoroughly checked for offensive material. If you find anything that you believe should be removed, please email the creators and let us know. We will consider whether to drop the image and whether to report the original image to Photobucket.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

The dataset does not identify subpopulations.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

It is possible to indirectly identify publishers and subjects using reverse image search

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

Images may contain information that allows people to make inferences about race, ethnicity, sexual orientation, religious beliefs, political opinions, memberships, locations, health information, or criminal history. However, because these images were shared publicly, we assume that that information is not considered too private to be shared.

Any other comments?

No.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The images were crawled from Photobucket and automatically aligned and cropped using dlib. The individual images were published in Photobucket by their respective authors under either Creative Commons BY 2.0, Creative Commons BY-NC 2.0, Public Domain Mark 1.0, Public Domain CC0 1.0, or U.S. Government Works license. All of these licenses allow free use, redistribution, and adaptation for non-commercial purposes.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated? If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Images were collected using a custom crawler to limit data scraped to those including permissive Creative Commons Licences. Sample data available for download was sampled randomly with a visual check for offensive content and basic demographic representativeness.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

No humans were involved in data collection and the data is not labeled. Humans involved in developing, testing, and executing the script and preparing it for publication were full time, paid employees of AVID.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Data was collected in 2018. Some data has been deleted since then, none has been added.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No separate ethical review process was conducted.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.

Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Via a third party (Creative Commons/Photobucket)

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Individuals were not notified of the data collection. They were aware that the images were public.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

No

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, 7 as well as a link or other access point to the mechanism (if appropriate).

Consent was not provided, but individuals who are in the dataset can petition to have their images removed by contacting AVID.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Metadata was deleted.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Most of the raw data still exists on photobucket, but we did not save the metadata we deleted.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

No

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

This dataset was originally curated to create a Generative Adversarial Network using a style-based generator architecture. The GAN created improved the state of the art in terms of established quality metrics.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No such repository exists.

What (other) tasks could the dataset be used for?

We believe that, if labelled, this dataset could create a face detection algorithm for use on "in the wild" images.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset was crawled from Photobucket and inherits the bias of that platform. In particular, we believe that the dataset may not be representative of the population in terms of skin color, age, and gender. Even if it is representative of the population in which it is used, there may be too few examples of very dark skin tones for the algorithm to do a good job. We believe based on research and past industry failures that oversampling minority classes may be necessary to bring performance on those classes up to par.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Caution should be taken when using this dataset on its own. Consider supplementing minority groups and auditing performance for dark skin tones, older adults, and women.

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be available publicly on GitHub.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The data is available on GitHub. It does not have a DOI.

When will the dataset be distributed?

The data was distributed in November 2018

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this

license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

NA

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

The individual images were published on Photobucket by their respective authors under either Creative Commons BY 2.0, Creative Commons BY-NC 2.0, Public Domain Mark 1.0, Public Domain CC0 1.0, or U.S. Government Works license. All of these licenses allow free use, redistribution, and adaptation for non-commercial purposes.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

NA

Any other comments?

Maintenance

Who is supporting/hosting/maintaining the dataset? The project team at AVID.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Contact form on our website

Is there an erratum? If so, please provide a link or other access point.

NA

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Dataset will only be updated if subjects or publishers request that their images be taken down. A changelog will be produced on the GitHub but we will not require that users update any local copies.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Because this dataset is a subset of a public corpus, our deleting data would not improve subjects' privacy. If the original corpus is challenged or removed on the grounds of harm to subjects, the publication of this dataset will be reconsidered.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

We do not anticipate versioning this dataset.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We do not expect to host an extended version of this dataset, however, others are welcome to clone the distribution and add on to it. We are particularly interested in annotations to the existing data and would be happy to cooperate within reason to extending the dataset with additional data or annotations.

Any other comments?

Appendix D: Problem Statement

Face detection algorithm

A national chain jewelry store has found that thieves tend to be aware of security cameras mounted on the ceiling and plans to add eye level cameras in high-traffic stores. They plan to first implement face detection using data from concealed, eye-level cameras. This model will be deployed at each store. It will first be used to collect images of customers' faces. Images of faces from use will be used to improve the model so that it can detect faces in each store environment. Later, the model will be supplemented with customer files and incident reports in hopes of adding functionality. For example, management hopes that individual stores will be able to catch repeat offenders and identify customers later found to be casing stores for later thefts. One day, they may use the original model and all store data to see if they can identify suspicious behavior across stores.

Problem: **formulate a plan for how you'd build a model to detect the presence of a face and identify key features in stills from video** with the context of the above plan in mind.

Data: The data involves images of faces in different orientations and with a wide variety of background features and accessories.

Appendix E: Problem Statement

Face detection algorithm

A national chain jewelry store has found that thieves tend to be aware of security cameras mounted on the ceiling and plans to add eye level cameras in high-traffic stores. They plan to first implement face detection using data from concealed, eye-level cameras. This model will be deployed at each store. It will first be used to collect images of customers' faces. Images of faces from use will be used to improve the model so that it can detect faces in each store environment. Later, the model will be supplemented with customer files and incident reports in hopes of adding functionality. For example, management hopes that individual stores will be able to catch repeat offenders and identify customers later found to be casing stores for later thefts. One day, they may use the original model and all store data to see if they can identify suspicious behavior across stores.

Problem: formulate a plan for how you'd build a model to detect the presence of a face and identify key features in stills from video with the context of the above plan in mind.

Data: The data involves images of faces in different orientations and with a wide variety of background features and accessories.

Appendix F: Training Data Sample

Here are the first 10 images (Figure F.1), last 10 images F.2), and 10 from the middle (Figure F.3) of the training dataset.

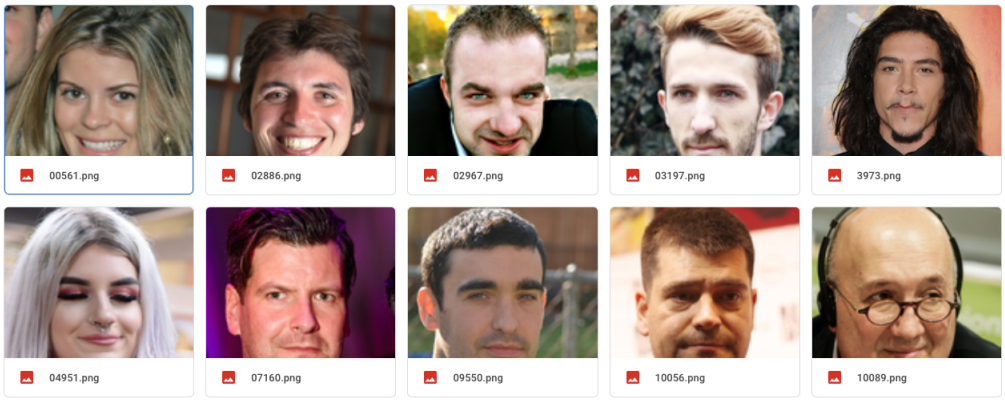


Figure F.1: First 10 images

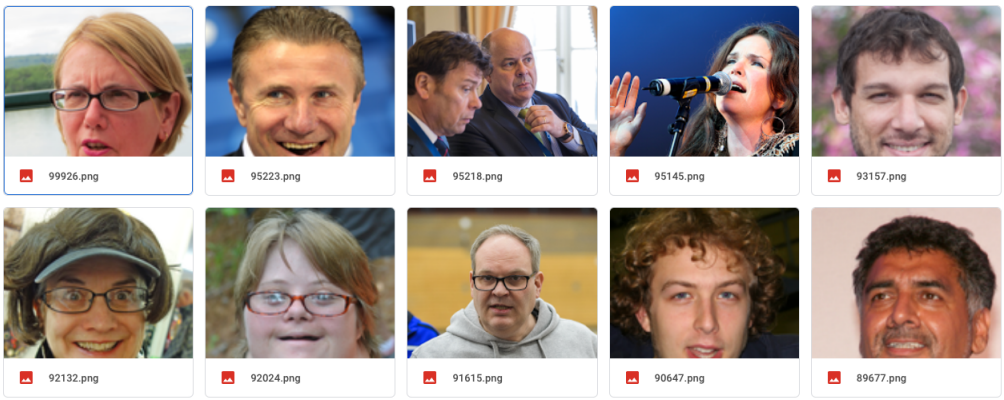


Figure F.2: Last 10 images

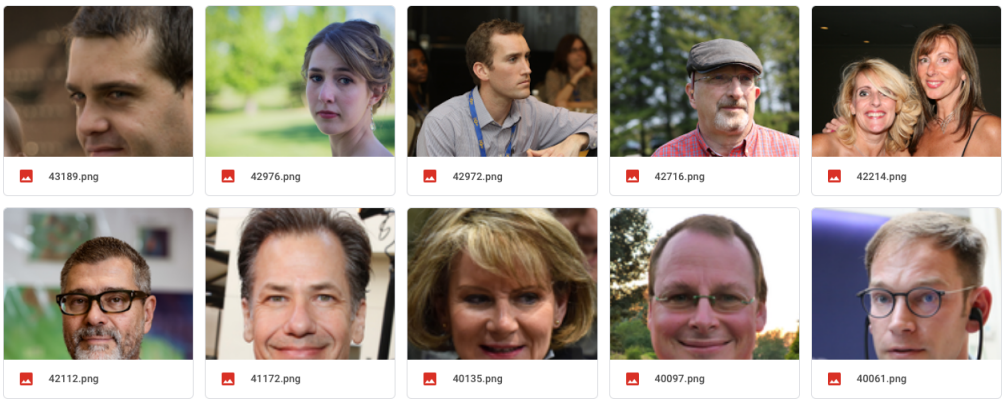


Figure F.3: Middle 10 images

Bibliography

[1] AI ethics toolkits - intel AI.

[2] AI fairness 360. Library Catalog: aif360.mybluemix.net.

[3] Beall's list of potential predatory journals and publishers.

[4] The hunt for mobile missiles: Nuclear weapons, AI, and the new arms race. Section: Special Reports.

[5] P7000 - IEEE draft model process for addressing ethical concerns during system design.

[6] Stack overflow developer survey 2019.

[7] Wisconsin v. loomis opinion.

[8] Philip Adler, Casey Falk, Sorelle A. Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. 54(1):95–122.

[9] Sung-Hee Ahn and Hye-A. Yeom. Moral sensitivity and critical thinking disposition of nursing students in korea. 20(5):482–489.

[10] Ishmael P. Akaah. Differences in research ethics judgments between male and female marketing professionals. 8(5):375–381.

[11] Ali A Al-Kazemi and Gary Zajac. Ethics sensitivity and awareness within organizations in kuwait: An empirical exploration of espoused theory and theory-in-use. page 9.

[12] Taghreed Alshehri, Reuben Kirkham, and Patrick Olivier. Scenario co-creation cards: A culturally sensitive tool for eliciting values. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–14. Association for Computing Machinery.

[13] Morgan G. Ames, Janet Go, Joseph 'Jofish' Kaye, and Mirjana Spasojevic. Understanding technology choices and values through social class. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, CSCW '11, pages 55–64. Association for Computing Machinery.

[14] Julia Angwin and Jeff Larson. Machine bias.

[15] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 167–176. Association for Computing Machinery.

[16] 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*.

[17] Solon Barocas and Andrew D. Selbst. Big data's disparate impact.

[18] Eric P.S. Baumer, Timothy Berrill, Sarah C. Botwinick, Jonathan L. Gonzales, Kevin Ho, Allison Kundrik, Luke Kwon, Tim LaRowe, Chanh P. Nguyen, Fredy Ramirez, Peter Schaedler, William Ulrich, Amber Wallace, Yuchen Wan, and Benjamin Weinfeld. What would you do?: Design fiction and ethics. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, pages 244–256. ACM.

[19] Muriel Bebeau, J Rest, and C Yamoor. Measuring dental students' ethical sensitivity. 49:225–35.

[20] Yahav Bechavod and Katrina Ligett. Learning fair classifiers: A regularization-inspired approach. abs/1707.00044.

[21] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. 6:587–604.

[22] Ruha Benjamin. Informed refusal: Toward a justice-based bioethics. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

[23] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations.

[24] Lawrence Blum. *Moral Perception and Particularity*.

[25] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

[26] Mary M. Brabeck, Lauren A. Rogers, Selcuk Sirin, Jennifer Henderson, Michael Benvenuto, Monica Weaver, and Kathleen Ting. Increasing ethical sensitivity to racial and gender intolerance in schools: Development of the.... 10(2):119.

[27] Stacy M. Branham, Anja Thieme, Lisa P. Nathan, Steve Harrison, Deborah Tatar, and Patrick Olivier. Co-creating & identity-making in CSCW: revisiting ethics in design research. In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW Companion '14*, pages 305–308. ACM Press.

[28] Barry Brown, Alexandra Weilenmann, Donald McMillan, and Airi Lampinen. Five provocations for ethical HCI research. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 852–863. Association for Computing Machinery.

[29] Amy Bruckman. Research ethics and HCI. In Judith S. Olson and Wendy A. Kellogg, editors, *Ways of Knowing in HCI*, pages 449–468. Springer.

[30] Louis L. Bucciarelli. *Designing engineers*. MIT Press.

[31] Howard Buchan. Reidenbach and robins multidimensional ethics scale: Testing a second-order factor model. 4(10).

[32] Joy Buolamwini. CODED BIAS.

[33] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classi64257;cation. page 15.

[34] Jenna Burrell. How the machine thinks: Understanding opacity in machine learning algorithms. 3(1):205395171562251.

[35] Kenneth D. Butterfield, Linda Klebe Trevin, and Gary R. Weaver. Moral awareness in business organizations: Influences of issue-related and social context factors. 53(7):981–1018.

[36] Lisa Marie Byrd. Development of an instrument to identify the virtues of expert nursing practice: byrds nurses ethical sensitivity test (byrds nest). page 186.

[37] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. 356(6334):183–186.

[38] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3992–4001. Curran Associates, Inc.

[39] Sarah Carr. AI gone mental: engagement and ethics in data-driven technology for mental health. 29(2):125–130. Publisher: Routledge _eprint: https://doi.org/10.1080/09638237.2020.1714011.

[40] Erin A. Cech. Culture of disengagement in engineering education? 39(1):42–72. Publisher: SAGE Publications Inc.

[41] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. Ranking with fairness constraints.

[42] Audrey Chia and Lim Swee Mee. The effects of issue characteristics on the recognition of moral issues. page 16.

[43] Shruthi Sai Chivukula, Colin M. Gray, and Jason A. Brier. Analyzing value discovery in design decisions through ethicography. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12. Association for Computing Machinery.

[44] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. Slice finder: Automated data slicing for model validation. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1550–1553. ISSN: 2375-026X.

[45] Henriikka Clarkeburn. A test for ethical sensitivity in science. 31(4):439–453.

[46] G. A. Claypool, D. F. Fetyko, and M. A. Pearson. Reactions to ethical dilemmas: A study pertaining to certified public accountants. 9(9):699–706.

[47] Mark Coeckelbergh. Regulation or responsibility? autonomy, moral imagination, and engineering.

[48] Michael P Coyne, Dawn W Massey, and Jay C Thibodeau. Raising students' ethical sensivity with a value relevance approach. 7.

[49] Kate Crawford and Jason Schultz. Big data and due process: Toward a framework to redress predictive privacy harms. 55:37.

[50] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women.

[51] Peggy Desautels. Gestalt shifts in moral perception. Library Catalog: www.semanticscholar.org.

[52] Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–14. ACM Press.

[53] Larry Dignan. Kronos, IBM partner on watson bot amid broader AI push for human resources.

[54] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73. ACM.

[55] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning.

[56] Benjamin H. Dotger. "i had no idea": Developing dispositional awareness and sensitivity through a cross-professional pedagogy. 26(4):805–812.

[57] Paul Dourish. *Where the Action Is: The Foundations of Embodied Interaction*. The MIT Press.

[58] Marina Drosou, H.v. Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. Diversity in big data: A review. 5(2):73–84. Publisher: Mary Ann Liebert, Inc., publishers.

[59] K. Anders Ericsson. The influence of experience and deliberate practice on the development of superior expert performance. In K. Anders Ericsson, Neil Charness, Paul J. Feltovich, and Robert R. Hoffman, editors, *The Cambridge Handbook of Expertise and Expert Performance*, pages 683–704. Cambridge University Press.

[60] K. Anders Ericsson and Herbert A. Simon. *Protocol analysis: Verbal reports as data*. Protocol analysis: Verbal reports as data. The MIT Press.

[61] Nermin Ersoy and Fügen Göz. The ethical sensitivity of nurses in turkey. 8(4):299–312.

[62] Wesley J. Erwin. Supervisor moral sensitivity. 40(2):115–127.

[63] Casey Fiesler. Ethical considerations for research involving (speculative) public data. 3:1–13.

[64] Casey Fiesler, Natalie Garrett, and Nathan Beard. What do we teach when we teach tech ethics?: A syllabi analysis. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pages 289–295. ACM.

[65] Erik Fisher, Michael O'Rourke, Robert Evans, Eric B. Kennedy, Michael E. Gorman, and Thomas P. Seager. Mapping the integrative field: taking stock of socio-technical collaborations. 2(1):39–61. Publisher: Routledge _eprint: https://doi.org/10.1080/23299460.2014.1001671.

[66] Kenneth R. Fleischmann, Russell Robbins, and William A. Wallace. Designing educational cases for intercultural information ethics: The importance of diversity, perspectives, values and pluralism. 50(1):4–14.

[67] Kenneth R. Fleischmann, William A. Wallace, and Justin M. Grimes. Computational modeling and human values: A comparative study of corporate, academic, and government research labs. In *2011 44th Hawaii International Conference on System Sciences (HICSS)*, pages 1 –10.

[68] Samantha Fowler, Dana Zeidler, and Troy Sadler. Moral sensitivity in the context of socioscientific issues in high school science students. 31(2):279–296.

[69] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness.

[70] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning.

[71] Batya Friedman. Value-sensitive design. 3(6):16–23.

[72] Batya Friedman and David Hendry. The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1145–1148. Association for Computing Machinery.

[73] William Gaver. Making spaces: how design workbooks work. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, page 1551. ACM Press.

[74] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for datasets.

[75] R. Stuart Geiger and David Ribes. Trace ethnography: Following coordination through documentary practices. In *2011 44th Hawaii International Conference on System Sciences*, pages 1–10. ISSN: 1530-1605.

[76] Jet Gispen. Ethics for designers.

[77] Colin Gray, Sai Shruthi Chivukula, and Ahreum Lee. What kind of work do "asshole designers" create? describing properties of ethical concern on reddit.

[78] Colin M. Gray. Revealing students ethical awareness during problem framing. 38(2):299–313. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jade.12190.

[79] Colin M. Gray and Shruthi Sai Chivukula. Ethical mediation in UX practice. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–11. Association for Computing Machinery.

[80] Colin M. Gray, Cesur Dagli, Muruvvet Demiral8208;Uzan, Funda Ergulec, Verily Tan, Abdullah A. Altuwaijri, Khendum Gyabak, Megan Hilligoss, Remzi Kizilboga, Kei Tomita, and Elizabeth Boling. Judgment and instructional design: How ID practitioners work in practice. 28(3):25–49. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/piq.21198.

[81] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. Accepted: 2019-01-03T00:00:45Z.

[82] Tristan Greene. GPT-3s bigotry is exactly why devs shouldnt use the internet to train AI.

[83] Jessica Guyunn. Google photos labeled black people 'gorillas'.

[84] Marc Hanheide, Moritz Göbelbecker, Graham S. Horn, Andrzej Pronobis, Kristoffer Sjöö, Alper Aydemir, Patric Jensfelt, Charles Gretton, Richard Dearden, Miroslav Janicek, Hendrik Zender, Geert-Jan Kruijff, Nick Hawes, and Jeremy L. Wyatt. Robot task planning and explanation in open and uncertain worlds. 247:119–150.

[85] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning.

[86] Carla L. Harenski, Olga Antonenko, Matthew S. Shane, and Kent A. Kiehl. Gender differences in neural mechanisms underlying moral sensitivity. 3(4):313–321. Publisher: Oxford Academic.

[87] Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization.

[88] P Hebert, E M Meslin, E V Dunn, N Byrne, and S R Reid. Evaluating ethical sensitivity in medical students: using vignettes as an instrument. 16(3):141–145.

[89] Anne Kari T Heggestad, Per Nortvedt, and Åshild Slettebø. The importance of moral sensitivity when including persons with dementia in qualitative research. 20(1):30–40.

[90] Jessica Hemberg and Elisabeth Bergdahl. Ethical sensitivity and perceptiveness in palliative home care through co-creation. 27(2):446–460.

[91] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards.

[92] David Hollingworth and Sean Valentine. The moderating effect of perceived organizational ethical context on employees ethical issue recognition and ethical judgments. 128(2):457–466.

[93] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need?

[94] Lara Houston, Steven J. Jackson, Daniela K. Rosner, Syed Ishtiaque Ahmed, Meg Young, and Laewoo Kang. Values in repair. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 1403–1414. Association for Computing Machinery.

[95] Hangyu Huang, Yun Ding, Honghong Wang, Kaveh Khoshnood, and Min Yang. The ethical sensitivity of health care professionals who care for patients living with HIV infection in hunan, china: A qualitative study. 29(2):266–274.

[96] Chuck Huff, Laura Barnard, and William Frey. Good computing: a pedagogically focused model of virtue in the practice of computing (part 1). 6(3):246–278.

[97] Lilly C. Irani and M. Six Silberman. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 611–620. Association for Computing Machinery.

[98] Lily Irani. Consortium for the Science of Sociotechnical Systems Research Summer Institute 2019.

[99] Steven J. Jackson, Tarleton Gillespie, and Sandy Payette. The policy knot: reintegrating policy, practice and design in cscw studies of social computing. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, CSCW '14, pages 588–602. Association for Computing Machinery.

[100] Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness.

[101] Nassim JafariNaimi, Lisa Nathan, and Ian Hargraves. Values as hypotheses: Design, inquiry, and the service of values. 31(4):91–104. Conference Name: Design Issues.

[102] Suzy Jagger. Ethical sensitivity: A foundation for moral judgment. 1:13–30.

[103] Irving Janis. Groupthink. In *A First Look at Communication Theory*, pages 235–246. McGrawHill.

[104] Mark Johnson. *Moral Imagination: Implications of Cognitive Science for Ethics*. University of Chicago Press.

[105] Thomas M. Jones. Ethical decision making by individuals in organizations: An issue-contingent model. 16(2):366–395. Publisher: Academy of Management.

[106] Jennifer Jordan. Taking the first step toward a moral action: A review of moral sensitivity measurement across domains. 168(3):323–359.

[107] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks.

[108] Niki Kilbertus, Adrià Gascón, Matt J. Kusner, Michael Veale, Krishna P. Gummadi, and Adrian Weller. Blind justice: Fairness with encrypted sensitive attributes.

[109] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 656–666. Curran Associates, Inc.

[110] Elizabeth Anne Kinsella. Practitioner reflection and judgement as phronesis. In Elizabeth Anne Kinsella and Allan Pitman, editors, *Phronesis as Professional Knowledge: Practical Wisdom in the Professions*, Professional Practice and Education: A Diversity of Voices, pages 35–52. SensePublishers.

[111] Hannah Laqueur and Ryan Copus. Machines learning justice: A new approach to the problems of inconsistency and bias in adjudication.

[112] Reed Larson and Mihaly Csikszentmihalyi. The experience sampling method. In Mihaly Csikszentmihalyi, editor, *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*, pages 21–34. Springer Netherlands.

[113] Rebecca Lind. Ethical sensitivity in viewer evaluations of a TV news investigative report. 23(4):535–561.

[114] Rebecca Ann Lind and David L. Rarick. Assessing ethical sensitivity in television news viewers: A preliminary investigation. 10(2):69–82.

[115] Rebecca Ann Lind and David L. Rarick. Viewer sensitivity to ethical issues in TV coverage of the clinton-flowers scandal. 16(2):169–181.

[116] Rebecca Ann Lind and Tammy Swenson-Lepper. Measuring sensitivity to conflicts of interest: A preliminary test of method. 19(1):43–62.

[117] Jie Liu, Binke Yuan, Yue-jia Luo, and Fang Cui. Intrinsic functional connectivity of medial prefrontal cortex predicts the individual moral bias in economic valuation partially through the moral sensitivity trait.

[118] Ewa Luger, Lachlan Urquhart, Tom Rodden, and Michael Golembewski. Playing the legal card: Using ideation cards to raise data protection issues within the design process. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 457–466. Association for Computing Machinery.

[119] Yotam Lurie and Shlomo Mark. Professional ethics of software engineers: An ethical framework. 22(2):417–434.

[120] Kim Lützén, Agneta Johansson, and Gun Nordström. Moral sensitivity: some differences between nurses and physicians. 7(6):520–530.

[121] Kim Lützén, Gun Nordström, and Mats Evertzon. Moral sensitivity in nursing practice. 9(3):131–138.

[122] Bradley Malin and Latanya Sweeney. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. 37(3):179–192.

[123] Nonna Martinov-Bennie and Rosina Mladenovic. Investigation of the impact of an ethical framework and an integrated ethics education on accounting students ethical sensitivity and judgment. 127(1):189–203.

[124] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. Auditing search engines for differential satisfaction across demographics. pages 626–633.

[125] Jacob Metcalf, Emanuel Moss, and Danah Boyd. Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics. 86(2):449–476. Publisher: Johns Hopkins University Press.

[126] Jessica K. Miller, Batya Friedman, Gavin Jancke, and Brian Gill. Value tensions in design: the value sensitive design, development, and appropriation of a corporation's groupware system. In *Proceedings of the 2007 international ACM conference on Supporting group work*, GROUP '07, pages 281–290. Association for Computing Machinery.

[127] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. pages 220–229.

[128] Pascal Molenberghs, Joshua Gapp, Bei Wang, Winnifred R. Louis, and Jean Decety. Increased moral sensitivity for outgroup perpetrators harming ingroup members. 26(1):225–233.

[129] Cosmin Munteanu, Heather Molyneaux, Wendy Moncur, Mario Romero, Susan O'Donnell, and John Vines. Situational ethics: Re-thinking approaches to formal ethics requirements for human-computer interaction. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 105–114. Association for Computing Machinery.

[130] Liisa Myyry and Klaus Helkama. The role of value priorities and professional ethics training in moral sensitivity. 31(1):35–50.

[131] Laura Nader. Up the anthropologist: Perspectives gained from studying up.

[132] Arvind Narayanan and Vitaly Shmatikov. How to break anonymity of the netflix prize dataset.

[133] Darcia F. Narvaez. Moral perception: A new construct?

[134] Helen Nissenbaum. How computer systems embody values. 34(3):120–119.

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

[136] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

[137] Denise M Patterson. Causal effects of regulatory, organizational and personal factors on ethical sensitivity. page 37.

[138] Lars Jacob Tynes Pedersen. See no evil: moral sensitivity in the formulation of business problems. 18(4):335–348.

[139] Robin R Radtke. The effects of gender and setting on accountants' ethically sensitive decisions. page 14.

[140] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. page 13.

[141] Akanksha Rana and Katie Paul. U.s. charges facebook with racial discrimination in targeted housing ads.

[142] Alison DeNisco Rayome. The US, china and the AI arms race: Cutting through the hype.

[143] R Eric Reidenbach and Donald P Robin. Some initial steps toward improving the measurement of ethical evaluations of marketing activities. pages 871–879.

[144] R Eric Reidenbach and Donald P Robin. Toward the development of a multidimensional scale for improving evaluations of business ethics. page 15.

[145] R Eric Reidenbach, Donald R Robin, and Lyndon Dawson. An application and extension of a multidimensional ethics scale to selected marketing practices and marketing groups. page 10.

[146] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. Algorithmic impact assessments. page 22.

[147] James Rest, Darcia Narvaez, Muriel J. Bebeau, and Stephen J. Thoma. *Postconventional moral thinking: A neo-Kohlbergian approach*. Postconventional moral thinking: A neo-Kohlbergian approach. Lawrence Erlbaum Associates Publishers. Pages: ix, 229.

[148] James R. Rest. A psychologist looks at the teaching of ethics. 12(1):29–36.

[149] Scott J Reynolds and Jared A Miller. The recognition of moral issues: moral awareness, moral sensitivity and moral attentiveness. 6:114–117.

[150] Diana Robertson, John Snarey, Opal Ousley, Keith Harenski, F. DuBois Bowman, Rick Gilkey, and Clinton Kilts. The neural processing of moral sensitivity to issues of justice and care. 45(4):755–766.

[151] Mostafa Roshanzadeh, Zohreh Vanaki, and Afsaneh Sadooghiasl. Sensitivity in ethical decision-making: The experiences of nurse managers. 27(5):1174–1186. Publisher: SAGE Publications Ltd.

[152] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. 1(5):206–215. Number: 5 Publisher: Nature Publishing Group.

[153] Troy D. Sadler. Moral sensitivity and its contribution to the resolution of socio8208;scientific issues. 33(3):339–358.

[154] Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. Integrating ethics within machine learning courses. 19(4):1–26.

[155] Steve Sawyer and Mohammad Hossein Jarrahi. Sociotechnical approaches to the study of information systems. pages 5–1–5–27. Publisher: CRC Press.

[156] Allen Schmaltz. On the utility of lay summaries and AI safety disclosures: Toward robust, open research oversight. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 1–6. Association for Computational Linguistics.

[157] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2503–2511. Curran Associates, Inc.

[158] Andrew D. Selbst. Disparate impact in big data policing.

[159] Michael K. Shaub. An empirical examination of the determinants of auditors' ethical sensitivity.

[160] Katie Shilton. Engaging values despite neutrality: Challenges and approaches to values reflection during the design of internet infrastructure. 43(2):247–269.

[161] Katie Shilton. Four billion little brothers?: Privacy, mobile phones, and ubiquitous data collection. 52(11):48–53.

[162] Katie Shilton. Values and ethics in human-computer interaction. 12(2):107–171.

[163] Katie Shilton. Values levers: Building ethics into design. 38(3):374–397.

[164] Katie Shilton and Daniel Greene. Linking platforms, practices, and developer ethics: Levers for privacy discourse in mobile application development. 155(1):131–146.

[165] Katie Shilton and Jes A. Koepfler. Making space for values: communication & values levers in a virtual team. In *Proceedings of the 6th International Conference on Communities and Technologies*, C&T '13, pages 110–119. Association for Computing Machinery.

[166] Ben Shneiderman. Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight. 113(48):13538–13540.

[167] Can Simga-Mugan, Bonita A. Daly, Dilek Onkal, and Lerzan Kavut. The influence of nationality and gender on ethical sensitivity: An application of the issue-contingent model. 57(2):139–159.

[168] John R. Sparks and Shelby D. Hunt. Marketing researcher ethical sensitivity: Conceptualization, measurement, and exploratory investigation. 62(2):92–109.

[169] Mudhakar Srivatsa and Mike Hicks. Deanonymizing mobility traces: Using social networks as a side-channel. page 10.

[170] Lucy Suchman, Jeanette Blomberg, Julian Orr, and Randall Trigg. Reconstructing technologies as social practice. 43(3):392–408. Num Pages: 17 Number: 3.

[171] Latanya Sweeney. Discrimination in online ad delivery.

[172] Tammy Swenson-Lepper. Ethical sensitivity for organizational communication issues: Examining individual and organizational differences. 59(3):205–231.

[173] Curtis Tate. Racial bias in facial recognition software: What travelers should know as TSA, CBP expand programs.

[174] Kristinn R. Thórisson. Integrated a.i. systems. 17(1):11–25.

[175] Linda Klebe Trevino. Ethical decision making in organizations: A person-situation interactionist model. 11(3):601–617. Publisher: Academy of Management.

[176] Linda K. Treviño, Gary R. Weaver, and Scott J. Reynolds. Behavioral ethics in organizations: A review. 32(6):951–990.

[177] Bruce W. Tuckman and Mary Ann C. Jensen. Stages of small-group development revisited. 2(4):419–427. Publisher: SAGE Publications.

[178] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. 15(2):49–60.

[179] Peter-Paul Verbeek. Materializing morality. 31(3):361 –380.

[180] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. ISSN: 1063-6919.

[181] Jessica Vitak, Katie Shilton, and Zahra Ashktorab. Beyond the belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 941–953. Association for Computing Machinery.

[182] Indrè Žliobaitè. A survey on measuring indirect discrimination in machine learning.

[183] Indrè Žliobaitè and Bart Custers. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. 24(2):183–201.

[184] Kathryn Weaver. Ethical sensitivity: State of knowledge and needs for further research. 14(2):141–155.

[185] Kathryn Weaver and Carl Mitcham. Prospects for developing ethical sensitivity in nursing, engineering, and other technical professions education. pages 1–18.

[186] Kathryn Weaver, Janice Morse, and Carl Mitcham. Ethical sensitivity in professional practice: concept analysis. 62(5):607–618.

[187] Kathryn Weaver and Janice M. Morse. Pragmatic utility: using analytical questions to explore the concept of ethical sensitivity. 20(3):191–214.

[188] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann. Google-Books-ID: 1SylCgAAQBAJ.

[189] Dennis Wittmer. Ethical sensitivity and managerial decisionmaking: An experiment. 2(4):443–462.

[190] Dennis P. Wittmer. Ethical sensitivity in management decisions: Developing and testing a perceptual measure among management and professional student groups. 4(2):181–205.

[191] Richmond Y Wong and Nick Merrill. Engaging speculative practices to probe values & ethics in sociotechnical systems. page 4.

[192] Richmond Y. Wong, Deirdre K. Mulligan, Ellen Van Wyk, James Pierce, and John Chuang. Eliciting values reflections by engaging privacy futures using design workbooks. 1:111:1–111:26.

[193] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, H. V. Jagadish, and Gerome Miklau. A nutritional label for rankings. pages 1773–1776.

[194] Scott A Yetmar and Kenneth K Eastman. Tax practitioners' ethical sensitivity: A model and empirical examination. page 18.

[195] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM.

[196] Doug Zytko, Jessa Lingel, Jeremy Birnholtz, Nicole B. Ellison, and Jeff Hancock. Online dating as pandora's box: Methodological issues for the CSCW community. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, CSCW'15 Companion, pages 131–134. Association for Computing Machinery.