# Datasheets for Datasets Help ML Engineers Notice and Understand Ethical Issues in Training Data

KAREN L. BOYD, University of Michigan, USA

The social computing community has demonstrated interest in the ethical issues sometimes produced by machine learning (ML) models, like violations of privacy, fairness, and accountability. This paper 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 asked to review an intentionally ethically problematic dataset and asked to think aloud as they used it to solve a given ML problem. Out of 23 participants, 11 were given a Datasheet they could use while completing the task. Participants were ethically sensitive enough to identify concerns 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. The think-aloud protocol offered a grounded description of how participants recognized, understood, and made a decision about ethical problems in an unfamiliar dataset. The method used in this study can test other interventions that claim to encourage recognition, promote understanding, and support decision-making among technologists.

CCS Concepts: • **Human-centered computing** → *Collaborative and social computing design and evaluation methods*; *User studies*; *Computer supported cooperative work*.

Additional Key Words and Phrases: training data, machine learning, ethics, ethical sensitivity, development practices

**ACM Reference Format:**
Karen L. Boyd. 2021. Datasheets for Datasets Help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 438 (October 2021), 27 pages. https://doi.org/10.1145/1122445.1122456

## 1 INTRODUCTION

Machine Learning (ML) finds patterns in training data, but does not distinguish between useful bias (that helps it differentiate between images of cats and cars, for example) and discriminatory bias (that may, for example, assess Black parolees as more likely to reoffend [32]). ML reflects bias in the world from which this data was drawn, for example in associations between words in text, resulting in algorithms that reify those biases [13]. Training data has also become a target for accountability interventions (e.g. [27, 24, 5]) and is described by industry practitioners as a key place to intervene to support fairness in ML [28]. Training data are an area of particular concern for privacy advocates, too, but the issue is complicated because collecting sensitive attributes (e.g. race or gender) may be necessary to build and certify fair algorithms [73]. Therefore, researchers advocate interventions into training data to preserve privacy and allow for fair model training,



Author's address: Karen L. Boyd, University of Michigan, Ann Arbor, Michigan, USA, karboyd@umich.edu.


2573-0142/2021/10-ART438 $15.00
https://doi.org/10.1145/1122445.1122456

model certification, and decision verification by, for example, encrypting sensitive attributes in training data [34]. Because it is cited as a driver of discrimination, a focus for those concerned with multiple ethical issues, and highlighted by industry practitioners as a target for intervention, this project focuses on a tool that aims to intervene to support ethical ML development while its builders are first working with training data.

Context documents are interventions 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 and understand ethical issues [5, 24, 58, 72, 50, 55]. As part of that effort, most 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 that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups?" [24]. 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" [5]. This paper offers some empirical data to support the idea that questions about dataset characteristics can prompt ethical engagement.

Ethical sensitivity (ES) gives us a framework to observe how context documents scaffold 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 understanding), and judgment (where the worker selects and executes a path forward). It has been studied in a variety of professions [68] and has been used to test educational interventions [46, 15, 30, 16].

This paper uses ES to describe how a selected context document– Datasheets for Datasets [24]– 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.

This paper answers the following research questions:

> **RQ 1**: How do machine learning engineers recognize, particularize, and make judgments about potential ethical problems in unfamiliar training data?
>
> **RQ 2**: How might Datasheets support ethical sensitivity among ML engineers working with unfamiliar training data?

More participants who were given Datasheets recognized ethical issues while working with the data and participants relied heavily on the Datasheet to particularize. Among participants who had a Datasheet, most particularization was done with the Datasheet on the screen.

Section 2 reviews current work on ethical sensitivity and ethical topics with training data. Section 3 describes methods, 4 discusses analysis, 5 explains the results, and 6 reviews the importance of these findings for ML development, ethical sensitivity, and ethical cooperative development.

## 2 LITERATURE REVIEW

This project contributes to an ongoing conversation in CSCW about describing, critically evaluating, and intervening in the work practices of technologists and designers [21, 70] in general and machine learning and data science practitioners [65, 45, 2, 14] in particular. To contextualize this study, 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 will introduce the category of interventions I call context documents and explain how I selected one to focus on. Finally, I will review the framework of Ethical Sensitivity and how context documents try to promote two of its components: ethical recognition and particularization.

### 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 will not 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 that are relevant when considering the training data used to build ML models. Fairness, privacy, and accountability are particularly relevant to the facial recognition dataset used in this study. Conceptions of all three vary among people and contexts, complicating the task of operationalizing them in technology.

*2.1.1 Fairness.* Training data can be a table of rows with features and a dependent variable, like a regression would need. Training data can also be images, video, free text, shopping histories, online learning activity, and more. Training data is biased. Statistical bias in training data is in fact what the algorithm uses to label an image as being of a cat and not a dog or a person or a car. But training data can also contain patterns that reflect real-world prejudice or systemic inequality, leading to discriminatory bias in the algorithm's outputs. An algorithm can't tell the difference between a pattern like "a longer nose-looking thing tends to correlate with the label 'dog' and a shallower face is more likely to be 'cat."' and the pattern "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: [48, 4]) and researchers have built useful taxonomies of discriminatory bias in machine learning [26, 42].

Discrimination in ML models not only comes from decisions reflecting societal prejudices, it can also manifest as performance differences among 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% [11].

In the ML fairness literature and among people impacted by algorithms, there are a wide range of definitions of fairness that can be sophisticated, contextually-determined, and contradictory with one another [25, 53, 59, 64]. Non-technical users of algorithms seem to be attuned to the possibility of bias in algorithms [38], and HCI research has developed methods of surfacing user values, perceptions of (un)fairness, and beliefs about what factors should and should not inform algorithmic decision-making [6, 12, 66]. ML engineers are in a unique position to evaluate and operationalize fairness concepts tailored for the users, subjects, and regulatory environment of algorithms.

*2.1.2 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 collected. Researchers have raised privacy concerns about this data [57] and about "notice and consent," the ethical safeguard

used for collecting it [3]. Research suggests that privacy is also complex and contextually-defined [47].

Including many, diverse examples in a training data set can address quality problems, including unfairness, but adding more data can mean more people's privacy is at risk. This is aggravated by the fact that to identify discriminatory bias, sensitive attributes may need to be collected and stored [73]. Interventions to address bias may prescribe oversampling rare or sensitive cases, meaning members of minority groups can be more likely to have data collected about them.

*2.1.3 Transparency.* The lack of visibility into most types machine learning models makes it difficult to identify potential for harm, locate the source of identified harm, or mitigate it. Some reasons for this lack of transparency are the technical restrictions on interpretability of the models themselves; the intellectual property protections, privacy laws, and organizational secrecy that safeguard the details of training data; the policies, people, and software that mediate algorithmic outputs; and sometimes a lack of disclosure or awareness that a model is in use at all. Many advocate for increased transparency into models' development process, data, inferences, deployment, potential harm, and human involvement in algorithms' development and deployment to allow its builders, users, citizens, courts, and regulators to understand, interpret, and act on their outputs [18, 43]. There's also discussion of the potential harms that total transparency could create [36]. Research has identified several particular areas of concern for algorithmic transparency, including calling for data provenance tools that can reflect data context and reuse practices [63], transparency practices that support decision-making and accountability [29], legal guidelines that account for quick technological change [56], and oversight bodies that can preserve privacy rights and manage perverse incentives while examining systems in close detail [36]. This paper focuses on a type of transparency intervention I will call "context documents" that attempt to expose relevant aspects of data or models to help builders and users make informed decisions.

## 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 such datasets [62]. Standard documentation accompanying datasets (or models) can bridge the gap between builders and users.

These "context documents" take many forms, ranging in complexity from a few hundred words [54] to detailed reports [5, 27]. Proposals like Bender and Friedman's [5] for Natural Language Processing and Yang et al.'s [72] 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 [50, 55, 58]. Gebru et al.'s Datasheets [24], Mitchell et al.'s Model Cards [44], and Yang et al.'s nutritional label [72] directly ask for information about ethical concerns, while others argue that simply reporting the characteristics of datasets will prompt and advertise ethical work [5, 27]. 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" [32], Bolukbasi et al.'s "Man is to Computer Programmer as Woman is to Homemaker?" [8], 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), for particular domains and techniques and for general use, and as deeply technical reports or lay language documents. Tables 1 and 2 classify each of the context documents mentioned above key dimensions.

Table 1. Context Documents by Scope and Focus

| | **Technique- or Domain-Specific** | **General Purpose** |
|---|---|---|
| **Training Data** | Bender & Friedman, 2018 (Natural Language Processing) | Holland et al., 2018; ***Gebru et al., 2018*** |
| **Model** | Diakopoulos, 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 2. Context Documents by Audience and Purpose

| **General Call/Program** | **Technical Reports** | **Lay Language Documents** |
|---|---|---|
| Shneiderman, 2016; Reisman et al., 2018; Diakopoulos, 2016 | Holland et al., 2018; Bender & Freidman, 2018; Yang et al., 2018 | Bender & Freidman, 2018; Schmaltz, 2018; Mitchell et al., 2019 ***Gebru et al., 2018*** |

This paper focuses on Datasheets [24]: 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. They are legible to key non-expert stakeholders, like managers, users, citizens, and auditors, and can therefore empower them and other interested parties to support accountability. Although Datasheets were built with a variety of goals [24], one of them is to increase the likelihood that ML engineers notice, understand, and can act on ethical problems in datasets.

### 2.3 Ethical sensitivity

Ethical sensitivity (ES) started as Rest's "moral sensitivity" [52] and has emerged as a way to understand how people recognize, interpret, and act on ethically consequential decisions in their work [68]. I argue that ES will help us understand how the people who build technology notice, understand, and act on ethical problems in their work; ES gives us a framework for evaluating ethical interventions into technology development; and that studies of technology development will add breadth and depth to current understandings of ES [9].

Although ethical sensitivity is often studied as a trait (e.g. by asking "are men or women more ethically sensitive?" [1, 49]) some studies suggest that it may be a skill that can be developed [19] or taught [22], or is a collection of related skills [39].

This paper 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 [41]), 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 [35].

According to this view of ES, a person may consciously practice the skill 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 ES strikes a chord with long-standing research models that assert that ethical decision-making is dependent on the worker and their work characteristics [61] and influenced by the particulars of the ethical issue [31].

ES is highly situated, but that does not mean we can not 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 as they practice. The practice session can be 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 with and without a Datasheet with data and a problem I control, so observing people in their full, long-term work context was not feasible. Therefore, I developed a think-aloud comparison method to allow ML engineers to work on a task I designed 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? Ethical sensitivity can be thought of as consisting of three activities: recognition, particularization and judgment.

*2.3.1 Recognition.* Recognition of an ethical issue is the moment of noticing. 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 [17, 49], giving the worker the opportunity to intervene.

There's little prior work in ethical sensitivity describing what Weaver et al. refer to as "cues" that trigger ethical recognition [68]. Context documents like Datasheets may operate as cues, and in fact, some are designed to do so. Papers proposing these documents talk about their allowing dataset or model users to "recognize . . . potential limitation" [5]. Holland et al. [27] 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. [44] noted that some "systematic errors were only exposed after models were put into use, and negatively affected users," hoping that Model Cards could help avoid these oversights.

*2.3.2 Particularization.* Particularization is a less well-defined and -studied area of ethical sensitivity [9]. Blum [7] explicates the importance of such particulars to ethical judgment. Weaver et al. [68] include activities that develop an understanding of the particulars of the ethical situation: reflecting on one's beliefs, seeking information about circumstances, and referring to external standards, like policies or codes of ethics. For the purposes of this study, particularization is any kind of understanding-building activity.

Context documents are not only created to spark recognition. Bender and Friedman [5] say that their report could "[allow] reasoning about what the likely effects may be." Mitchell et al. [44] 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 [54] 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 [24].

The process of building an understanding of the particulars of an ethically consequential situation is under-explored in the existing ES literature [9].

*2.3.3 Judgment.* Rest's foundational work lays out a clear conceptualization of judgment, identifying 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" [52]. In later work, Rest et al. acknowledged that sensitivity and judgment are intertwined: "Logically, Component 1 [what he called "moral sensitivity"] precedes Component 2 [what he called "moral 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 inter-actions" [51].

In machine learning development, judgment could lead an engineer to, for example, implement a tool to mitigate bias in training data, encrypt training data, refuse to build, or continue along the development plan with no adjustments. Although judgment is sometimes not included in conceptions of ES, its direct connection to design outcomes make it important to observe.

This study uses ethical sensitivity to evaluate the effectiveness of Datasheets, especially whether they help ML engineers recognize and particularize. The think-aloud method may offer insight into particularization that is not available with the survey methods used extensively in that literature. I designed this study to focus on particularization, but it also captures recognition and judgment among participants.

## 3 METHODS

This paper 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 dataset.

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. A randomly selected half of them were provided a Datasheet along with the problem statement and data, and participants each worked for 25 minutes or until they deemed themselves ready to describe a plan.

### 3.1 Participants

Participants were recruited through the Slack channel for an ML Meet-up group the author attends (6), referrals 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 in Table 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).

In order to avoid framing and anchoring effects in the interviews, none of the interview protocols explicitly mention race or gender, nor did I collect this demographic information about participants. Particularly, I was concerned that asking about these race and gender characteristics in a nuanced and considerate way might directly influence the study protocol and limit how participants were thinking. As described in 6.5, I believe this is an important direction of future research.

Table 3. Participants

| | | Datasheet | No Datasheet | Total |
|---|---|---|---|---|
| **Experience (years)** | Min | .25 | .1 | .1 |
| | Max | 15 | 5.5 | 15 |
| | Average | 5.0 | 2.1 | 3.5 |
| | Median | 3 | 2 | 2 |
| **Job Role** | Worker | 5 | 7 | 12 |
| | Student | 3 | 5 | 8 |
| | Manager | 2 | | 2 |
| | Volunteer | 1 | | 1 |
| **Industry** | Industry | 9 | 6 | 15 |
| | Academia | 1 | 1 | 2 |
| | Both | | 2 | 2 |
| | Unclear | 1 | 3 | 4 |

All participants consented to have the audio and screen-sharing recorded and one recording failed (Participant 21).

### 3.2 Think Aloud

I approached my research questions with a think aloud comparison study. 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 [20].

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.

Verbalizations that require minimal cognitive processing offer information are not thought to impede creativity, change decision-making, or alter the structure of task performance [20]. There is evidence to suggest that they slow down task performance, however, so recorded times (i.e. time spent looking at Datasheets) can 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 Materials

*3.3.1 Problem Statement.* I provided participants with a fictitious problem statement, provided here in its entirety:

> "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."

This problem was selected because it has a variety of ethical issues that could be noticed and 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 for non-Black participants.

*3.3.2 Data.* I selected 171 images from the Flickr-Faces-HQ (FFHQ) dataset [33]. FFHQ includes faces "in the wild" pulled from the photo sharing site Flickr and is well documented. To manufacture demographic imbalance that could cause an ethical issue, I intentionally oversampled images of people who appeared to be men and who were light-skinned. This data was presented by the script as "a random sample" from a larger dataset.

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 race was not clear. 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's Syndrome and two subjects with dramatic costume make-up. Various ages were represented, including young children.

*3.3.3 Datasheet.* The Datasheet intervention was designed by Gebru et al. [24][2]. I filled out the Datasheet with information from the original dataset's curators [3] and added fictionalized details to suit the purposes of the study. The goals of fictionalizing were to present enough details that participants could find and act on real details about the images' source if they executed a web search, but to ensure that they did not believe that the extensive documentation about the FFHQ dataset reflected the composition of the intentionally-biased study dataset. Fictionalized details included replacing Flickr with Photobucket as the source. Photobucket is a similar site for which user demographics are similar to those of the fake dataset and are readily available to any participant

[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.

[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

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 was not associated with the experiment because it could muddy participant interpretations of data provenance and could cause participants to confuse this intentionally under-representative dataset with the original.

The Datasheet acknowledged two potential ethical issues explicitly. First, the data reflects the demographics of its source, which is heavily male and white. 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. The other ethical issues planted in the problem statement were not acknowledged in the Datasheet.

### 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. I asked them to think aloud as they decided "whether and how" to use the data for 25 minutes. Their screen and audio were recorded (all participants consented. One recording failed: P21DS). Avoiding interrupting participants as they spoke, participants were asked to stop working after around about 25 minutes. Several participants naturally concluded earlier, offering a summary of their next steps, and a few reflected and searched for longer.

After the think-aloud session, I asked questions using a funnel-sequence interview [40]. Inspired by [60], 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 was not 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 clearly described the three time categories used in funnel sequence interviews [60]: during Time A, a participant describes their perception, during Time B participants are asked about moral aspects of the situation without being directly asked about ethics, and during 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).

## 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 the text. 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. 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. I also labeled any recognition, particularization and judgment that occurred in the post-think-aloud interview as being "prompted" ethical sensitivity verbalizations.

### 4.1 Recognition

The first time a participant mentioned a particular ethical issue 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]?" However, participants struggled to respond to this question, and gave answers about the issue itself, not their noticing. 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. For example, participants might mention a facial recognition dataset or model that had been in the news for its ethical complications without voicing an objection (I did not count this as recognition) or describe facial recognition as "scary" (I did count this as recognition). I decided to count an instance as recognition only if participants outwardly expressed concern, even if their concern was vague. Fortunately, all the participants with ambiguous utterances later exhibited clear, unprompted recognition. This means that these instances did not change the count of unprompted recognition, but did mean I did not consider time-to-first-recognition to be a meaningful measure. This ambiguity could be a weakness in this way of measuring recognition or it could indicate that recognition is not always a single awakening moment, but can also be a dawning realization.

### 4.2 Particularization

For the purpose of identifying particularization, I developed the following definition, based primarily on Weaver et al. [67, 68] Blum [7] 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."

While coding transcripts, I labeled these utterances as "particularization," and recorded the time, notes about context, and screen contents. If the participant was seeking or citing remembered information, I recorded the information's source and topic.

I recorded which options participants considered, constraints they mentioned, examples or stories they offered, comments about signals of credibility in sources they referred to, and comments that signaled the participants' understanding of the technical and social aspects of the ethical issue (how it happens and its impacts).

### 4.3 Judgment

Although this study was not aimed at collecting judgments, some participants offered judgments as they worked.

I labeled a statement as a judgment if 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").

## 5 RESULTS

Participants demonstrated ethical sensitivity, with and without Datasheets. Although participants in the group who were given Datasheets had more unprompted recognition, suggesting that the Datasheet may have aided with recognition, all the participants in the group with no Datasheets recognized at least one ethical issue after being prompted by interview questions. Ten out of eleven participants who were given a Datasheet read it, and participants given Datasheets relied on it extensively while they were building an understanding of the ethical issue at hand.

The method offered insight into recognition, particularization and judgment among machine learning engineers using unfamiliar training data for a simulated work task. The difficulty of locating recognition in a single moment of "awakening," which the literature led me to expect [68] 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. Particularization revealed how they relied on their existing social and technical understandings, which existing understanding and novel facts they believed to be relevant, and how they synthesized those to make sense of the problem, their options, their resources, and the risks of different courses of action. Although judgment wasn't a target of the method used, it revealed the diversity of interventions participants considered, their specific instrumental goals, and how their socio-technical understanding informed their decision-making.

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

### 5.1 Recognition

Figure 1 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. Table 4 shows what each participant had on their screen when they noticed their first ethical issue.

Out of 11 participants who received a Datasheet, 10 read at least some of the document. One participant in the Datasheets group did not read the Datasheet and did not mention any ethical issues, including in response to direct ethics question.

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 that photographers published the photos and often subjects' consent was unknown. P21DS highlighted this question in the document, saying, "So the subjects didn't give permission?"

P1DS and P17DS mentioned bias while reading about the data selection.

> "So random selection of those, good. . . 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

Table 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 | | |

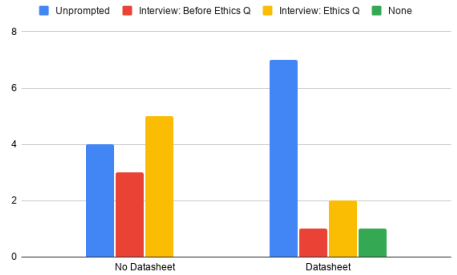


Fig. 1. Prompted and unprompted issue recognition

> skin tones. So kind of leads to racist machine learning more or less." *-P17DS (User highlighted the phrase "basic demographic representativeness")*

One participant in the group without Datasheets and one participant in the group with Datasheets first mentioned an ethical issue while looking at the data.

Participants mentioned 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).

### 5.2 Particularization

Participants who particularized out loud as they worked spent a widely variable amount of time doing so, ranging from 33 seconds to 9 minutes and 17 seconds (average: 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 specific ethical issues they recognized primarily. In fact, they built and reflected on broader technical and social understanding. Particularization utterances revealed that participants' understandings differed substantially.

While developing and exploring both types of understanding, participants reflected on their existing knowledge and beliefs, sought (or indicated they planned to seek) information, and relied on the study materials. They used examples of past engineering failures (8) and successes (1) 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.

*5.2.1 Social Understanding.* While deciding what to do, participants considered the perspectives of and relied on their beliefs about other people, including data subjects, shoppers, thieves, the business implementing the system, and law enforcement. The example of law enforcement shows how differing views about stakeholders shapes participants' view of the morality of the system.

P11DS used their relationship to law enforcement to support their moral evaluation of the project.

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

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

> "For certain, especially ethnic groups . . 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. . . 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. How do you even make that dataset of 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 individuals more than others." *-P5N*

*5.2.2 Technical Understanding.* Participants relied on and sought information about the data, the technology, and deployment circumstances. Participants were particularly interested in certain aspects of the data: its source, its distribution, how it was curated, how it was sampled or selected from the source, whether and how it was tested, and what changes had been made to it. Source and distribution information were mentioned as ways to identify bias. When citing their own technical understanding, there were areas of consensus and conflict.

Participants differed in their views of some essential training data composition questions including whether 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 talked 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*

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 [reading Datasheet] oh, 'is centered on the images center pixel.' Okay, so that's as I feared. . . For an actual production situation, we have to deal with not just detecting a face, but also centering things." *-P17DS*

In contrast to the disagreement about negative examples, there was widespread agreement that differences between training data and the production scenario would be a problem. 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?. . . 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 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 for that, like I said, one would need labeled data. . . And 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." *-P18N*

*5.2.3 Problem, Options, Resources, and Risks.* To understand a problem, its sources, effects, and what can be done about it, participants synthesized social and technical understandings.

P3DS relied on understandings of related technology and users' perspectives to understand the ethical aspects of technical choices.

> "So if I make a new commit [to a GitHub repository] where I got rid of John's awkward photo. . . [it] 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. GitHub is probably not where you want to store this if you want to be able to have a revised data set." *-P3DS*

> "People still don't necessarily assume that everybody is going to be able to access this stuff even when they make it public. Some people don't realize the default settings. . . So they might have not realized they were opting in to us collecting their information." *-P3DS*

Just as there were differences among social and technical understandings, differences emerged when addressing socio-technical issues, like false positives.

P20N considers the costs of false positives ethically and financially.

> "You know, this isn't holding a kid while you call their mom . . . This is people's lives are potentially on the line. And is that worth saving $500 for a stolen ring? . . . 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 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, is using a model here going to provide benefit or not? And how can we 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, the goal is to catch as much as possible. So, not a problem if it's false. . . 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*

*5.2.4 Datasheets and Particularization.* In the group with Datasheets, most particularization happened while the Datasheet was on the screen. 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 about Creative Commons licenses, P3DS felt somewhat assured, but it didn't resolve the concern entirely: "At least legally, we look like we're okay even if whether we're okay, morally slash ethically– might not be the same question."

P3DS used information from the Datasheet to reason about provenance and bias:

> "Probably ask whoever sent me this how they determined [that each image includes a human face]? 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? . . . are you getting some sort of bias here in terms of, you only have images that have easily recognizable faces because something already recognized that there was a face in."

> "It might be nice if they said why it was deleted to see– any time 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."

*Participant 11.* P11DS initially focused very directly on technical aspects of the task. Although they nearly instantly acknowledged the possibility of an ethical issue (51 seconds in) and read the entire Datasheet, 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, [the author] will take care of it, don't care."

However, after reading more, the participant reported a paradigm shift, not unlike what's described in the ES literature as an awakening [67].

> "[*Reading*]'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]. . . 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 started with very different senses of the relationship between the ethical issues and their task, both demonstrated high ethical sensitivity (unprompted recognition and particularization), and both used the Datasheet to shape their perception of the ethical issue.

*5.2.5 Particularizing without a Datasheet.* Four participants without Datasheets recognized during the think-aloud, 3 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:

> "You can try to train a network to do anything you want. But there has to be sort of pattern. 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– there's really bias and everything. So there will definitely be bias in your training data. . . 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), mostly reflecting on the circumstances of use: the behavior of thieves and innocent shoppers, the setting of stores in malls, mall security, and the relatively diverse demographics of the U.S. Notably, P10N mentioned the "Coded Bias" project [10] 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, including the behavior of innocent shoppers and thieves, and compared 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. . . 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."

### 5.3 Judgment

As I expected, participants did not do as much judging during the short think-aloud session as they did particularizing, but some interesting data about judgment emerged. 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.

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.

The most popular solution mentioned for dealing with bias was altering the demographic distribution of the training data. However, proposed ideas differed on two dimensions: the goal and the means. Participants often changed their minds here and their ideas did not align with others'.

Participants goals included altering the demographic distribution to reflect the jewelry store locations (P4DS, P13DS, P14N, P17DS); ensure the data is equally distributed across groups (P8N, P10N, P15N, P16N); match U.S. demographics (P8N, P17DS), match criminal population (P5N), or in such a way that accuracy is similar among groups (P3DS, P13DS). To accomplish their goals, participants considered collecting more data (P2N, P3DS, P4DS, P7DS, P9DS, P13DS, P19N, P20N), reweighing (P4DS, P13DS, P15N, P18N), over-sampling minority groups (P3DS, P11DS, P16N, P18N,

P19N), under-sampling majority groups (P16N, P18N), unspecified data augmentation (P9DS, P14N, P18N), artificially generating more data for minority groups (P10N, P9DS), darkening existing images of light-skinned people (P6N, P11DS), deleting images for which the algorithm does not work well (P7DS), using a fairness toolkit (P9DS), and doing more testing (P18N, P9DS, P15N). P10N mentioned twice that they wanted to spend more time collecting data and feature engineering and went on to explain why, revealing a connection between their judgment and socio-technical understanding:

> "And I have recently– This is a really big issue, the 'Coded Bias' was the one thing, the recent article that I read . . in Stanford or MIT, the– it was able to detect more of white people than black people. . . because the training data consisted more of white people. So I think . . we should spend more time collecting data and you know, build the model, because model building is– I think with the architecture that we have now, with the computing power it's not as difficult . . . the collection of data is the one thing which we lack these days"

P11DS changed their mind as they reflected on a strategy to adjust the distribution, 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. . . 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*

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 a non-linear relationship between particularization and judgment [51].

## 6 DISCUSSION

This study contributes to the ongoing conversation about contextualized ethical behavior and offers a method for observing recognition, particularization, and judgment among technologists. It offers support and guidance for developing interventions into technologists' practices, ideas for the development and testing of curricula and policy, rich contextualized data about ethical sensitivity, and suggests opportunities for useful future work.

### 6.1 Context Documents and Ethical Sensitivity

This study suggests that context documents in general and Datasheets in particular may 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 encourage ethical sensitivity. More participants in this study with Datasheets mentioned ethical issues while working than those who did not. Participants relied on them extensively to particularize: most particularization in the Datasheets group happened while looking at the Datasheet. Although it's tough to evaluate ethical judgment without declaring some judgments better than others, four participants suggested the drastic step of replacing the problematic 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 a better measure of whether a participant thought it was relevant during the think-aloud than whether

they noticed it. If the Datasheet has that effect in the workplace – signaling to ML engineers that data context and ethical aspects are relevant to their work and thereby encouraging them to bring it up– it is still achieving its aims.

When I proposed this study, I was concerned that participants might not read the Datasheet: I had a plan to re-balance the groups 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 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 6 recognition events, 4 of which were the participant's first. 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 may be correct in believing that surfacing information about data distribution and context may trigger recognition, even without including direct ethical questions or language [5]. The fact that 4 participants (1 with a Datasheet and 3 without) mentioned their first ethical issue in response to early, indirect interview questions further supports the assertion that surfacing dataset characteristics and likely effects may prompt ethical engagement.

### 6.2 Fostering Ethical Sensitivity

I hope that further study of ethical sensitivity among technologists will offer a foundation for an evidence-based exploration of interventions to develop ethical sensitivity. For example, how can organizations embed ethical sensitivity in norms and policies? How can training develop the skill of ethical sensitivity in new hires and students? Besides context documents, organizations and educators can consider developing other tools, practices, and policies or shaping norms to encourage recognition, support particularization, and guide judgment. This study offers some guidance for developing and evaluating these interventions.

As we learn more about ethical sensitivity among technologists, we can build curricula that support ethical sensitivity. Recognition skills may be supported by teaching familiarity with common cues, an understanding of relevant ethical issues, historical examples of when and how new social impacts have emerged, and a habit of actively looking for ethical problems. Students may improve particularization skills by learning to evaluate information sources, building familiarity with existing options and resources, methods for evaluating risks, understanding important aspects of fit between problems and interventions, and developing a socio-technical perspective of their work and its context. Judgment may be supported through practice making or evaluating judgments in case studies and by providing preparation for and experience navigating the (perhaps unexpected) loop between particularization and judgment.

Participants in this study had a variety of educational backgrounds. Ideally, curricula should be adaptable for the several ways ML engineers can be trained, including through online courses outside the university, more distributed self-teaching, and on-the-job training. Finally, ethical sensitivity has been operationalized to test the success of educational interventions in professional education [15, 16, 46] and we can extend and adapt that work to study ML learners.

Organizations can develop and test policies, norms, and training to support the development of ethical sensitivity. Once asked about ethical issues in the interview, all but one participant cited at least one concern. This suggests that an intervention that involves simply asking technologists about potential ethical issues may go a long way. Maybe better than consistent questions as part of

a regular meeting or form, which could eventually prompt a habitual "no," are intermittent prompts: perhaps something analogous to the experience sampling method [37].

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 workers' responsibility and rewarding that engagement in job reviews with positive feedback, raises, and promotions could go a long way to ensuring that engineers are looking for and are willing to report potential ethical issues.

### 6.3 Think Aloud & Ethical Sensitivity

This study applied the think-aloud method to observe ethical sensitivity among technologists working with unfamiliar training data. Think-aloud has difficulties and advantages, but overall renders a rich view of ES compared to traditional methods of observing ethical sensitivity.

When applied to studying ethical sensitivity, the think-aloud method really shines when it comes to observing particularization. Until now, studies of particularization have been inconsistent and acontextual, like asking participants to rank or list factors that they considered when responding to a scenario. Think aloud, even in a simulated work context, allowed me to watch participants search for information, use examples, rely on existing understanding, and reflect. It revealed what existing understanding mattered and how those understandings differed among participants. Think aloud will give ES researchers a more grounded and more complete conceptualization of particularization and insight into how context effects the process.

Think-aloud may also offer 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.

Despite Rest's description of a non-linear relationship between particularization and judgment [51], the operationalization of ES used in studies up to this point has not left room for describing, let alone contending with, this complicated relationship [9]. Participants in this is study demonstrated this pattern, moving from judgment back to particularization and using their updated understanding in their next judgment phase. The number, triggers, and qualities of these interruptions of judgment to return to particularization could be useful data. Does fewer loops back to particularization reflect a confident, experienced worker? Perhaps more frequent or longer pauses to particularize are desirable for a sensitive, fraught, or complicated socio-technical project. More qualitatively, what causes and characterizes these loops? What makes them useful, and when are they simply delays?

This method was able to measure the quantity of recognition events, identify cues for recognition, and observe recognition as it unfolded. The data in this study suggests that recognition may not always look like a "moment of awakening" or a sudden paradigm shift, but perhaps a response to

an accumulation of concern. If that's the case, perhaps a wide-net approach of many interventions throughout the development process, will be most effective. Future studies have the opportunity to focus on this key moment so we can learn more about it and build best practices to identify or elicit details about it. Better understanding recognition presents an opportunity to improve the effectiveness of ethics practices in organizations because noticing an ethical issue is necessary to open the door for other practice-based and technical interventions for ethical ML.

*6.3.1 Situational Factors.* When observing ethics practices among technologists, researchers have a lot of options for how to design or select a work task and environment. These choices have trade-offs and offer insight into different aspects of the phenomenon of interest.

First, how will you deal with individual familiarity? My participants varied in their familiarity with image data, facial recognition technology, the regulatory environment of facial recognition and retail, the particular ethical issues embedded in the task, and experience with ML in general. This study dealt with these differences through randomization, but other studies could study a problem particular to a certain domain (or data type, ML technique, regulatory environment, or social context) with engineers with deep domain experience to reveal more about domain-specific problems and how expertise mediates ethical sensitivity.

Researchers will also have to manage the obviousness of any problems or cues they embed in a designed task. Some factors that may effect the obviousness are news attention to the ethical issue, social media discourse about it, or regulatory changes. Pilot testing can help researchers get a sense of how salient any planted cues are to people who are taking them in along with a lot of other new information. Manipulating how obvious a cue or problem is will allow researchers to focus on different aspects of ethical sensitivity. For example a study with a very large number of participants and a subtle ethical problem can offer a more nuanced view of recognition, while a more top-of-mind issue like the one selected for this study can offer more data from later stages.

How much time will participants spend working on your task? Although just 30 minutes offered a lot of data in this study, observing participants over the course of a longer project would offer insight into important features of real work that are not clear in the first minutes of working with new data. In particular, it would reveal how particularization and judgment continue to feedback on each other over time, how new information or testing alter their understanding, and whether and how their considerations of options, resources, and risks change over time as they make design commitments. Kaggle competitions are a fairly constrained, but much longer term project that engineers are often quite invested in but do not risk their employers' intellectual property.

Finally, researchers must make decisions about the environment in which participants will work on the simulated project. Participants can be observed in their real work environment; on their own computer or a provided computer; in their own environment or a lab environment. Lab environments and provided computers offer control, more granular data collection without requiring participants to download invasive software, and higher comparability among participants. This could be useful for attempting to measure things like time-to-first-recognition or collecting granular behavior data like keyboard and mouse input or eye-tracking. Participants' own environment or real work environment does not require participants to use unfamiliar software and settings or spend time configuring it. Unfamiliar software, settings, and surroundings may serve as recurring signals of the task's distance from their occupation– when we are interested in part in their way of thinking and working within their occupation, this can be a disadvantage.

### 6.4 Limitations

This study demonstrated that we can observe recognition and test the effectiveness of an intervention designed to prompt it using a think aloud experiment. The ethical issue selected in this study

has attracted a lot of attention in the news, including during the study period: participants mentioned the Tiny Images dataset being taken offline by MIT and the relevance of Black Lives Matter protests in the wake of the killing of George Floyd, which were ongoing during data collection. It is no surprise that most participants noticed ethical issues in facial recognition, especially when paired with the goal of detecting crime.

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 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.

In my small sample, randomization didn't render an even distribution of the demographics I did collect, particularly experience. In addition to designing studies focusing on the relationship between experience and ethical sensitivity, future researchers may consider making an effort to balance experience among study groups if experience is an explanatory variable.

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 work contexts as possible to ensure that we get an accurate picture.

### 6.5 Future Work

This study suggests several profitable avenues for future work.

We can learn much more about Datasheets' effectiveness as an ethical intervention. This study aimed to observe recognition in detail and to capture some particularization, and so an issue that was expected to be frequently noticed was selected. To further evaluate the Datasheet's effectiveness, it should be tested with more ambiguous or less talked-about ethical issues as well. More targeted work on particularization could quantify time spent on different sections of the Datasheet which, in conjunction with verbalizations about the content, could give us an even better view of how Datasheets support particularization and how we might improve them. Recent work on Algorithmic Impact Assessments (AIAs) supports the importance of Datasheets and other context documents for supporting AIAs, but also suggests that filling them out may not be as simple as it seems [43]. Future studies should develop and test best practices for writing effective, thorough Datasheets that consider the lived realities of affected communities.

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 would be interesting see whether interventions like envisioning cards [23], adversary cards[69], and design workbooks [71] elicit or change the character of ethical sensitivity.

All 3 participants without Datasheets particularized for longer than average. It may be that a Datasheet supports more efficient particularization but with only 3 non-Datasheet particularizers, this study does not offer enough data to be sure. 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.

Participants' awareness of and beliefs about particular ethical issues were not a focus of this study, but they very likely influence particularization and judgment. Data about participants' positionality– like race, gender, sexuality, country of origin, and more– were not collected for this study and also represent an opportunity for future work. Experience with the particular technical features of the assigned task, application domain, or ethical issues could impact whether participants exhibit ethical sensitivity in a simulated work task, as well as their confidence and clarity in expressing their concerns. Perhaps job role, years of experience, or industry may play a role. Future work can help us understand how ethical awareness, beliefs, positionality, and experience influence or mediate ethical sensitivity during work will allow practitioners to appropriately target interventions in team composition, worker training, and task design to support ethical sensitivity.

## 7 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 paper offers a view into ethical sensitivity in technology development and reports the most detailed, contextual description of particularization and judgment yet. It offers evidence supporting Rest's assertion about the relationship between particularization and judgment in practice and raises questions about the current view of recognition.

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 paper 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.

## 8 ACKNOWLEDGMENTS

I would like to thank my advisor, Katie Shilton, for her guidance and encouragement throughout graduate school and on this project. Thanks to the members of my dissertation committee, including Katie, Hal Daumé III, Wayne Lutters, Yla Tausczik, Jessica Vitak, and Susan Winter, who helped me develop the dissertation of which this paper was part. I would also like to thank the anonymous referees for their valuable comments on this paper. This work is supported by the National Science Foundation under grant number 1452854.

## REFERENCES

[1] Ishmael P. Akaah. "Differences in research ethics judgments between male and female marketing professionals". In: *Journal of Business Ethics* 8.5 (May 1989), pp. 375–381. ISSN: 0167-4544, 1573-0697. DOI: 10.1007/BF00381729. URL: http://link.springer.com/10.1007/BF00381729 (visited on 12/07/2019).

[2] Cecilia Aragon et al. "Developing a Research Agenda for Human-Centered Data Science". In: *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. CSCW '16 Companion. New York, NY, USA: Association for Computing Machinery, Feb. 2016, pp. 529–535. ISBN: 978-1-4503-3950-6. DOI: 10.1145/2818052.2855518. URL: https://doi.org/10.1145/2818052.2855518 (visited on 04/03/2021).

[3] Solon Barocas and Helen Nissenbaum. "On notice: The trouble with Notice and Consent". In: *Proceedings of the Engaging Data Forum: The First International Forum on the Application and Management of Personal Electronic Information*. 2009. URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2567409 (visited on 06/24/2016).

[4] Solon Barocas and Andrew D. Selbst. "Big Data's Disparate Impact". In: *SSRN Electronic Journal* (2016). ISSN: 1556-5068. DOI: 10.2139/ssrn.2477899. URL: https://www.ssrn.com/abstract=2477899 (visited on 02/27/2019).

[5] Emily M. Bender and Batya Friedman. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science". In: *Transactions of the Association for Computational Linguistics* 6 (Dec. 2018), pp. 587–604. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00041. URL: https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00041 (visited on 02/28/2019).

[6] Niels van Berkel et al. "Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study". en. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019), pp. 1–21. ISSN: 2573-0142. DOI: 10.1145/3359130. URL: https://dl.acm.org/doi/10.1145/3359130 (visited on 03/27/2021).

[7] Lawrence Blum. *Moral Perception and Particularity*. 1991. URL: www.jstor.org/stable/2381661 (visited on 12/02/2019).

[8] Tolga Bolukbasi et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 4349–4357. URL: http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf (visited on 01/08/2019).

[9] Karen Boyd and Katie Shilton. "Adapting Ethical Sensitivity as a Construct to Study Technology Design Teams (forthcoming)". In: *Proceedings of the 2022 ACM Conference on Supporting Groupwork* (2022).

[10] Joy Buolamwini. *CODED BIAS*. CODED BIAS. URL: https://www.codedbias.com (visited on 09/29/2020).

[11] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (2018), p. 15.

[12] Jenna Burrell et al. "When Users Control the Algorithms: Values Expressed in Practices on Twitter". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019), 138:1–138:20. DOI: 10.1145/3359240. URL: https://doi.org/10.1145/3359240 (visited on 04/03/2021).

[13] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (Apr. 14, 2017), pp. 183–186. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aal4230. arXiv: 1608.07187. URL: http://arxiv.org/abs/1608.07187 (visited on 01/08/2019).

[14] Stevie Chancellor, Eric P. S. Baumer, and Munmun De Choudhury. "Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019), 147:1–147:32. DOI: 10.1145/3359249. URL: https://doi.org/10.1145/3359249 (visited on 04/03/2021).

[15] Henriikka Clarkeburn. "A Test for Ethical Sensitivity in Science". In: *Journal of Moral Education* 31.4 (Dec. 2002), pp. 439–453. ISSN: 03057240. DOI: 10.1080/0305724022000029662. URL: http://search.ebscohost.com/login.aspx?direct=true&db=asn&AN=8633678&site=ehost-live (visited on 12/13/2019).

[16] Michael P Coyne, Dawn W Massey, and Jay C Thibodeau. "Raising Students' Ethical Sensivity with a Value Relevance Approach". In: *Advances in Accounting Education: teaching and Curriculum Innovation* 7 (2005).

[17] Peggy Desautels. "Gestalt Shifts in Moral Perception". In: *Mind and Morals: Essays on Cognitive Science and Ethics*. Library Catalog: www.semanticscholar.org. 1996, pp. 129–143.

[18] Nicholas Diakopoulos. "Algorithmic Accountability". In: *Digital Journalism* 3.3 (May 2015), pp. 398–415. ISSN: 2167-0811. DOI: 10.1080/21670811.2014.976411. URL: https://doi.org/10.1080/21670811.2014.976411 (visited on 02/28/2019).

[19] Benjamin H. Dotger. ""I Had No Idea": Developing Dispositional Awareness and Sensitivity through a Cross-Professional Pedagogy". In: *Teaching and Teacher Education: An International Journal of Research and Studies* 26.4 (May 2010), pp. 805–812. ISSN: 0742-051X. DOI: 10.1016/j.tate.2009.10.017. (Visited on 03/06/2020).

[20] K. Anders Ericsson and Herbert A. Simon. *Protocol analysis: Verbal reports as data*. Protocol analysis: Verbal reports as data. Cambridge, MA, US: The MIT Press, 1984. 426 pp. ISBN: 978-0-262-55012-3.

[21] Melanie Feinberg. "Material Vision". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. New York, NY, USA: Association for Computing Machinery, Feb. 2017, pp. 604–617. ISBN: 978-1-4503-4335-0. DOI: 10.1145/2998181.2998204. URL: https://doi.org/10.1145/2998181.2998204 (visited on 04/03/2021).

[22] Samantha Fowler, Dana Zeidler, and Troy Sadler. "Moral Sensitivity in the Context of Socioscientific Issues in High School Science Students". In: *International Journal of Science Education* 31.2 (2009), pp. 279–296.

[23] 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. Austin, Texas, USA: Association for Computing Machinery, May 5, 2012, pp. 1145–1148. ISBN: 978-1-4503-1015-4. DOI: 10.1145/2207676.2208562. URL: http://doi.org/10.1145/2207676.2208562 (visited on 08/06/2020).

[24] Timnit Gebru et al. "Datasheets for Datasets". In: *arXiv:1803.09010 [cs]* (Mar. 23, 2018). arXiv: 1803.09010. URL: http://arxiv.org/abs/1803.09010 (visited on 02/28/2019).

[25] Nina Grgic-Hlaca et al. "Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction". In: *Proceedings of the 2018 World Wide Web Conference*. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, Apr. 2018, pp. 903–912. ISBN: 978-1-4503-5639-8. DOI: 10.1145/3178876.3186138. URL: https://doi.org/10.1145/3178876.3186138 (visited on 04/03/2021).

[26] Thomas Hellström, Virginia Dignum, and Suna Bensch. "Bias in Machine Learning – What is it Good for?" In: *arXiv:2004.00686 [cs]* (Sept. 2020). arXiv: 2004.00686. URL: http://arxiv.org/abs/2004.00686 (visited on 07/02/2021).

[27] Sarah Holland et al. "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards". In: *arXiv:1805.03677 [cs]* (May 9, 2018). arXiv: 1805.03677. URL: http://arxiv.org/abs/1805.03677 (visited on 02/28/2019).

[28] Kenneth Holstein et al. "Improving fairness in machine learning systems: What do industry practitioners need?" In: *arXiv:1812.05239 [cs]* (Dec. 12, 2018). DOI: 10.1145/3290605.3300830. arXiv: 1812.05239. URL: http://arxiv.org/abs/1812.05239 (visited on 03/14/2019).

[29] Ben Hutchinson et al. "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure". en. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada: ACM, Mar. 2021, pp. 560–575. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445918. URL: https://dl.acm.org/doi/10.1145/3442188.3445918 (visited on 07/02/2021).

[30] Suzy Jagger. "Ethical sensitivity: A Foundation for Moral Judgment". In: *Journal of Business Ethics Education* 1 (Jan. 1, 2011), pp. 13–30. DOI: 10.5840/jbee2011813.

[31] Thomas M. Jones. "Ethical Decision Making by Individuals in Organizations: An Issue-Contingent Model". In: *The Academy of Management Review* 16.2 (Apr. 1991), p. 366. ISSN: 03637425. DOI: 10.2307/258867. URL: http://www.jstor.org/stable/258867?origin=crossref (visited on 05/31/2018).

[32] Jeff Larson Julia Angwin. *Machine Bias*. ProPublica. May 23, 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (visited on 04/13/2018).

[33] Tero Karras, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *arXiv:1812.04948 [cs, stat]* (Mar. 29, 2019). arXiv: 1812.04948. URL: http://arxiv.org/abs/1812.04948 (visited on 07/24/2020).

[34] Niki Kilbertus et al. "Blind Justice: Fairness with Encrypted Sensitive Attributes". In: *arXiv:1806.03281 [cs, stat]* (June 8, 2018). arXiv: 1806.03281. URL: http://arxiv.org/abs/1806.03281 (visited on 01/12/2019).

[35] Elizabeth Anne Kinsella. "Practitioner Reflection and Judgement as Phronesis". In: *Phronesis as Professional Knowledge: Practical Wisdom in the Professions*. Ed. by Elizabeth Anne Kinsella and Allan Pitman. Professional Practice and Education: A Diversity of Voices. Rotterdam: SensePublishers, 2012, pp. 35–52. ISBN: 978-94-6091-731-8. DOI: 10.1007/978-94-6091-731-8_3. URL: https://doi.org/10.1007/978-94-6091-731-8_3 (visited on 09/28/2020).

[36] Paul B. de Laat. "Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?" en. In: *Philosophy & Technology* 31.4 (Dec. 2018), pp. 525–541. ISSN: 2210-5433, 2210-5441. DOI: 10.1007/s13347-017-0293-z. URL: http://link.springer.com/10.1007/s13347-017-0293-z (visited on 07/02/2021).

[37] Reed Larson and Mihaly Csikszentmihalyi. "The Experience Sampling Method". In: *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*. Ed. by Mihaly Csikszentmihalyi. Dordrecht: Springer Netherlands, 2014, pp. 21–34. ISBN: 978-94-017-9088-8. DOI: 10.1007/978-94-017-9088-8_2. URL: https://doi.org/10.1007/978-94-017-9088-8_2 (visited on 09/29/2020).

[38] Min Kyung Lee and Su Baykal. "Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division". en. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Portland Oregon USA: ACM, Feb. 2017, pp. 1035–1048. ISBN: 978-1-4503-4335-0. DOI: 10.1145/2998181.2998230. URL: https://dl.acm.org/doi/10.1145/2998181.2998230 (visited on 03/27/2021).

[39] Rebecca Lind. "Ethical Sensitivity in Viewer Evaluations of a TV News Investigative Report". In: *Human Communication Research* 23.4 (June 1997), pp. 535–561. ISSN: 03603989. DOI: 10.1111/j.1468-2958.1997.tb00409.x. URL: https://academic.oup.com/hcr/article/23/4/535-561/4564970 (visited on 04/09/2020).

[40] Rebecca Ann Lind and Tammy Swenson-Lepper. "Measuring Sensitivity to Conflicts of Interest: A Preliminary Test of Method". In: *Science and Engineering Ethics* 19.1 (Mar. 1, 2013), pp. 43–62. ISSN: 1471-5546. DOI: 10.1007/s11948-011-9319-6. URL: https://doi.org/10.1007/s11948-011-9319-6 (visited on 02/07/2020).

[41] Kim Lützén, Gun Nordström, and Mats Evertzon. "Moral Sensitivity in Nursing Practice". In: *Scandinavian Journal of Caring Sciences* 9.3 (1995), pp. 131–138. ISSN: 1471-6712. DOI: 10.1111/j.1471-6712.1995.tb00403.x. URL: http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1471-6712.1995.tb00403.x (visited on 12/12/2019).

[42] Ninareh Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning". In: *arXiv:1908.09635 [cs]* (Sept. 2019). arXiv: 1908.09635. URL: http://arxiv.org/abs/1908.09635 (visited on 07/02/2021).

[43] Jacob Metcalf et al. "Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA:

Association for Computing Machinery, Mar. 2021, pp. 735–746. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445935. URL: https://doi.org/10.1145/3442188.3445935 (visited on 04/03/2021).

[44] Margaret Mitchell et al. "Model Cards for Model Reporting". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), pp. 220–229. DOI: 10.1145/3287560.3287596. arXiv: 1810.03993. URL: http://arxiv.org/abs/1810.03993 (visited on 03/27/2019).

[45] Michael Muller et al. "Interrogating Data Science". In: *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. CSCW '20 Companion. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 467–473. ISBN: 978-1-4503-8059-1. DOI: 10.1145/3406865.3418584. URL: https://doi.org/10.1145/3406865.3418584 (visited on 04/03/2021).

[46] Liisa Myyry and Klaus Helkama. "The Role of Value Priorities and Professional Ethics Training in Moral Sensitivity". In: *Journal of Moral Education* 31.1 (Mar. 1, 2002), pp. 35–50. ISSN: 0305-7240. DOI: 10.1080/03057240120111427. URL: https://doi.org/10.1080/03057240120111427 (visited on 02/27/2020).

[47] Helen Nissenbaum. "Privacy as Contextual Integrity". In: *Washington Law Review* 79 (2004), p. 119. URL: https://heinonline.org/HOL/Page?handle=hein.journals/washlr79&id=129&div=&collection=.

[48] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016. 274 pp. ISBN: 978-0-553-41881-1.

[49] Robin R Radtke. "The Effects of Gender and Setting on Accountants' Ethically Sensitive Decisions". en. In: *Journal of Business Ethics* 24.4 (2000), pp. 299–312.

[50] Dillon Reisman et al. *Algorithmic impact assessments: A practical framework for public agency accountability*. en. Tech. rep. AI Now Institute, 2018.

[51] James Rest et al. *Postconventional moral thinking: A neo-Kohlbergian approach*. Postconventional moral thinking: A neo-Kohlbergian approach. Pages: ix, 229. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 1999. ix, 229. ISBN: 978-0-8058-3285-3.

[52] James R. Rest. "A Psychologist Looks at the Teaching of Ethics". In: *The Hastings Center Report* 12.1 (1982), pp. 29–36. ISSN: 0093-0334. DOI: 10.2307/3560621. URL: http://www.jstor.org/stable/3560621 (visited on 07/23/2019).

[53] Nripsuta Ani Saxena et al. "How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19. New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 99–106. ISBN: 978-1-4503-6324-2. DOI: 10.1145/3306618.3314248. URL: https://doi.org/10.1145/3306618.3314248 (visited on 04/03/2021).

[54] 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*. New Orleans, Louisiana, USA: Association for Computational Linguistics, June 2018, pp. 1–6. URL: http://www.aclweb.org/anthology/W18-0801 (visited on 02/28/2019).

[55] Andrew D. Selbst. "Disparate Impact in Big Data Policing". In: *SSRN Electronic Journal* (2017). ISSN: 1556-5068. DOI: 10.2139/ssrn.2819182. URL: https://www.ssrn.com/abstract=2819182 (visited on 02/28/2019).

[56] Hetan Shah. "Algorithmic accountability". en. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2128 (Sept. 2018), p. 20170362. ISSN: 1364-503X, 1471-2962. DOI: 10.1098/rsta.2017.0362. URL: https://royalsocietypublishing.org/doi/10.1098/rsta.2017.0362 (visited on 07/02/2021).

[57] Katie Shilton. "Four billion little brothers?: Privacy, mobile phones, and ubiquitous data collection". In: *Communications of the ACM* 52.11 (2009), pp. 48–53. URL: http://dl.acm.org/citation.cfm?id=1592778 (visited on 08/18/2016).

[58] Ben Shneiderman. "Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight". In: *Proceedings of the National Academy of Sciences* 113.48 (Nov. 29, 2016), pp. 13538–13540. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1618211113. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1618211113 (visited on 02/28/2019).

[59] Megha Srivastava, Hoda Heidari, and Andreas Krause. "Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. New York, NY, USA: Association for Computing Machinery, July 2019, pp. 2459–2468. ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330664. URL: https://doi.org/10.1145/3292500.3330664 (visited on 04/03/2021).

[60] Tammy Swenson-Lepper. "Ethical Sensitivity for Organizational Communication Issues: Examining Individual and Organizational Differences". In: *Journal of Business Ethics* 59.3 (July 1, 2005), pp. 205–231. ISSN: 1573-0697. DOI: 10.1007/s10551-005-2925-y. URL: https://doi.org/10.1007/s10551-005-2925-y (visited on 12/06/2019).

[61] Linda Klebe Trevino et al. "Managing Ethics and Legal Compliance: What Works And What Hurts". In: *California Management Review* 41.2 (Jan. 1, 1999), pp. 131–151. ISSN: 0008-1256, 2162-8564. DOI: 10.2307/41165990. URL: http://cmr.ucpress.edu/content/41/2/131 (visited on 04/10/2017).

[62] Joaquin Vanschoren. *OpenML*. OpenML: exploring machine learning better, together. URL: https://www.openml.org (visited on 04/15/2019).

[63] Michael Veale, Max Van Kleek, and Reuben Binns. "Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making". en. In: (2018), p. 14.

[64] S. Verma and J. Rubin. "Fairness Definitions Explained". In: *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. May 2018, pp. 1–7. DOI: 10.23919/FAIRWARE.2018.8452913.

[65] 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. New York, NY, USA: Association for Computing Machinery, Feb. 2016, pp. 941–953. ISBN: 978-1-4503-3592-8. DOI: 10.1145/2818048.2820078. URL: https://doi.org/10.1145/2818048.2820078 (visited on 04/03/2021).

[66] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. "Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–14. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376813. URL: https://doi.org/10.1145/3313831.3376813 (visited on 04/03/2021).

[67] Kathryn Weaver. "Ethical Sensitivity: State of Knowledge and Needs for Further Research". In: *Nursing Ethics* 14.2 (Mar. 2007), pp. 141–155. ISSN: 0969-7330, 1477-0989. DOI: 10.1177/0969733007073694. URL: http://journals.sagepub.com/doi/10.1177/0969733007073694 (visited on 09/09/2019).

[68] Kathryn Weaver, Janice Morse, and Carl Mitcham. "Ethical sensitivity in professional practice: concept analysis". In: *Journal of Advanced Nursing* (2008). URL: https://onlinelibrary-wiley-com.proxy-um.researchport.umd.edu/doi/full/10.1111/j.1365-2648.2008.04625.x (visited on 09/18/2019).

[69] Richmond Y Wong and Nick Merrill. "Engaging Speculative Practices to Probe Values & Ethics in Sociotechnical Systems". In: *Workshop at iConference*. 2019, p. 4.

[70] Richmond Y. Wong et al. "Beyond Checklist Approaches to Ethics in Design". In: *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. CSCW '20 Companion. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 511–517. ISBN: 978-1-4503-8059-1. DOI: 10.1145/3406865.3418590. URL: https://doi.org/10.1145/3406865.3418590 (visited on 04/03/2021).

[71] Richmond Y. Wong et al. "Eliciting Values Reflections by Engaging Privacy Futures Using Design Workbooks". In: *Proceedings of the ACM on Human-Computer Interaction* 1 (CSCW Dec. 6, 2017), 111:1–111:26. DOI: 10.1145/3134746. URL: http://doi.org/10.1145/3134746 (visited on 08/06/2020).

[72] Ke Yang et al. "A Nutritional Label for Rankings". In: *Proceedings of the 2018 International Conference on Management of Data - SIGMOD '18* (2018), pp. 1773–1776. DOI: 10.1145/3183713.3193568. arXiv: 1804.07890. URL: http://arxiv.org/abs/1804.07890 (visited on 02/28/2019).

[73] Žliobaitė, Indrė and Custers, Bart. "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models". In: *Artificial Intelligence and Law* 24.2 (June 1, 2016), pp. 183–201. ISSN: 1572-8382. DOI: 10.1007/s10506-016-9182-5. URL: https://doi.org/10.1007/s10506-016-9182-5 (visited on 10/08/2020).

Received January 2021; revised April 2021; revised July 2021; accepted July 2021