Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development

Karen L. Boyd
karenlboyd@gmail.com
San Diego Workforce Partnership
San Diego, California, USA

ABSTRACT

If “studying up,” or researching powerful actors in a social system, can offer insight into the workings and effects of power in social systems, this paper argues that “designing up” will give researchers and designers a tool to intervene. This paper offers a conception of “designing up,” applies the structure of Value Sensitive Design (VSD) to accomplish it, and submits an example of a tool designed to support ethical sensitivity, especially particularization and judgment. The designed artifact is a field guide for ethical mitigation strategies that uses tool profiles and filters to aid machine learning (ML) engineers as they build understanding of an ethical issue they have recognized and as they match the particulars of their problem to a technical ethical mitigation. This guide may broaden its users’ awareness of potential ethical issues, important features of ethical issues and their mitigations, and the breadth of available mitigations. Additionally, it may encourage ethical sensitivity in future ML projects. Feedback from ML engineers and technology ethics researchers rendered several usability improvements and ideas for future development. The tool can be found at: https://ml-ethics-tool.web.app/.

CCS CONCEPTS

• **Human-centered computing** → *Collaborative and social computing design and evaluation methods*; *Computer supported cooperative work*.

KEYWORDS

datasets, machine learning, ethics, ethical sensitivity, AI ethics, development practices

ACM Reference Format:
Karen L. Boyd. 2022. Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3534626

Machine learning (ML)-driven software is often built to support one party as they make decisions about others. Power differentials are common between the decision-maker and decision-subject, for example in medicine (including genomic, diagnostic, and mental health data [16, 19, 48]), law enforcement (including crime prediction and parole evaluations [2, 31]), employment (including hiring and evaluation [49]) and credit [49]. As AI systems proliferate, existing power relationships over who can get on an airplane [22], who gets laid off or promoted [49], and other decisions we allow one party to make about another will be informed, mediated, and legitimated by technology. Unfair, opaque, and invasive machine learning (ML) sometimes results from (e.g., [56]), and other times simply reifies (e.g., [16]) existing power dynamics in social systems.

Researchers and designers have developed interventions into training data, algorithms, or results to reduce bias, improve accountability and transparency, and reduce privacy risks (e.g.[28, 38, 41, 63].) Governments, organizations, researchers, and advocates have designed policies, ML techniques, educational campaigns, and other technologies to be used by a variety of actors to reduce harm, liability, and to protect these human values for their own sake (e.g.[25, 28, 32, 53]).

The diversity of these interventions designed to mitigate harm in ML is evidence of a promising attention to ethics throughout the ML ecosystems and at many moments along the development pipeline, targeting a variety of people and using different definitions of and legitimations for the values they aim to protect. Empowering citizens and data subjects to try to protect themselves with education, browser plug-ins, and the ability to opt out is important, but leaving it up to individuals isn’t enough: it puts pressure on the people with the least power in that circumstance; requires the education, coordination, and action of a very large group of people; and is limited in its scope of effect. A wide-net, “yes, and” approach to ML ethics will improve the chances that we catch and mitigate any given novel threat.

For example, in order to meaningfully disrupt the larger practice of surreptitious, extensive data collection and the use of that data to build ML algorithms that serve ads in a sometimes biased and harmful way [58], a huge number of users would have to install a browser plug-in. Designing a browser to block tracking by default is one way of intervening upstream: one design decision affects the privacy of many people. Alternatively, a designer could build tools for algorithm developers, supporting decisions to circumvent or reduce the harm of pervasive tracking and biased advertising for users, recognizing that design practices can support ethical reflection and intervention [57]. This paper proposes and demonstrates a method for designing for high-power actors in the social system surrounding ML, far upstream from users, enabling and encouraging them to mitigate risks for low-power actors. This is "designing up;" rather than designing a tool that can be used by lower-power

FAccT ’22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3534626

people to protect themselves, it's designing a tool that can be used by high-power, upstream actors to make choices that protect many. The designed tool can be found at https://ml-ethics-tool.web.app/

I employ Value Sensitive Design [29] to the project of designing up for machine learning engineers: explicitly defending the values of lower-power actors while designing a field guide to support machine learning engineers as they build an understanding of a potential ethical issue in their work and decide whether and how to act on it. I used ethical sensitivity [11] to operationalize the activities of building understanding as "particularization" and deciding how to act as "judgment." Using conceptual and empirical investigations, I identified key aspects of fit between ethical problems and mitigation tools. Then, I designed a search tool that allows engineers to search using these key aspects of fit. I hope this tool will make it easier and more efficient for ML engineers to understand and address ethical problems in their work and introduce ML engineers, educators, students, managers, and researchers to the broad range of ethical ML research and design. The tool can be filtered by key aspects of the ethical problem and its technical context; each mitigation strategy has a profile that describes in more detail key features and links to content sought by engineers; and users can participate and expand the project by suggesting edits, submitting tool profiles, or forking the project. It targets key harms created or propagated by ML– including privacy threats, outcome unfairness, procedural unfairness, lack of diversity, and lack of transparency– and intervenes with high-power actors who are upstream enough to mitigate harm.

1 BACKGROUND

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

1.1 Designing Up

Laura Nader encouraged anthropologists to not only study groups that are lower power in a social system, but also those in middle- and high-power positions. She notes that people with high power in social systems have broad public impact and responsibility [47]. If we replace her phrases relating to "studying" with "designing" and apply the discussion of people with high institutional power to those who build, sell, and use algorithms, we can repurpose some of her motivation as a call to action for designers in the age of proliferating and powerful algorithms:

> "the quality of life and our lives themselves my depend on the extent to which [designers intervene with] those who shape [socially-consequential algorithms] and actually control [their design, implementation, and use]. The [practice of design] is confronted with an unprecedented situation– never before have so few, by their actions and inactions, had the power of live and death over so many members of the species."

Nader says that studying those in power in social systems allows us to "flip" our questions: "Instead of asking why some people are poor, we would ask why other people are so affluent?" This allows us to understand and critique power in social systems. Studying up has had a significant impact on the field of anthropology [33] and has informed the study of technologists [55]. Designing up is not a method, but an orientation. It is not opposed, but complementary to other methods, movements, and values: a person can design up for accessibility, fairness, transparency, intersectional justice, or trans rights. They can use any of a variety of design methods to do so. It is also not a replacement for, but a complement to other design orientations: we want to recruit, but not rely on powerful actors to change the world for the better.

In FAccT, Barabas et al. argued for an analogous reorientation in data science [3]. In their case study, they executed a similar flip: rather than studying re-offense risk of prospective parolees (a project noted for its racial bias [2]), they focus on judges and judicial culture. They argue that data scientists who study up "could lay the foundation for more robust forms of accountability and deeper understandings of the structural factors that produce undesirable social outcomes via algorithmic systems."

This project echoes the call for designing up *using* data science, and extends it to designing *for* data science development, answering the call for supports for ethical algorithm design that are integrated into technologists' workflows and adaptable to organizational and industry contexts [21, 44]. Inspired by Irani's encouragement to use design to intervene "upstream" from harm [39, 40], I will use Value Sensitive Design (VSD) [29] to design a tool for data scientists and ML engineers to make it easier for them to employ ethical mitigations. VSD includes empirical investigations, which allow us to *study* up– with the aim of better understanding the workings of power in the design of influential technologies– and technical investigations that will let us *design* up– with the goal of intervening in the early stages of ML development.

1.2 Ethical Sensitivity

To reduce harms to fairness, privacy, and accountability from ML algorithms, this project aims to help ML engineers understand the particulars of and make decisions about the potential ethical problems in their work.

To operationalize these goals, I use the ethical sensitivity (ES) framework [11, 62]. ES describes a worker, focused on the technical aspects of their task, who experiences a paradigm shift (*recognition*) when they realize their task may be ethically consequential. They reflect and seek information about their situation: the particulars of the circumstances, opinions of relevant actors, stakeholder interests, relevant internal or external standards, resources, options, consequences of their options, and the relationship between the potential ethical issue and their own responsibilities (this stage is called *particularization*). Using what they learned, they formulate, select, and execute a *judgment*. Ethical sensitivity has been used to understand and intervene in many professions for decades [11, 62], and has recently been used to describe the ethical behavior or ML engineers [12]. FAccT has demonstrated interest in ethical reflection among technology's designers and builders [6]; ethical sensitivity offers an option for conceptualizing and operationalizing it. This study extends the work on ethical sensitivity among technologists and uses ethical sensitivity as a guide for design.

This project focuses on the second and third activities: after a worker recognizes and ethical issue and begins the search for a mitigation tool, their goal may be direct (to make a judgment) but in order to select an effective mitigation tool, they must have an understanding of (at least) the features of the ethical problem and candidate mitigation tools. These details, and the appropriateness of the mitigation tool features for the problem features, are what I'll refer to as "fit" throughout this paper. This project investigates worker needs: what features of fit matter when evaluating options? What features are nice to have, but not essential? What elements of work context matter when seeking fit?

Particularization and judgment are not linear stages– a workers often make developmental judgments, seek more information, and re-evaluate [51]. However, analytically separating information seeking and reflection from the decisions they support will help us understand where framing of interventions should aim to help engineers with building understanding (e.g., informative messaging) or decision-making (e.g., persuasive messaging).

1.2.1 *Particularization.*

Past work has observed ML engineers as they reflect and seek information about many types of "particulars" [12]. Just as in other industries [62], particularization is broad. It can include external information and internal beliefs about features of the circumstance, the stakeholders, the ethical issue(s), options, resources, and consequences. This design project aims to directly support a key activity within particularization: seeking information about options.

Of course, types of particulars are interrelated. For example, in order to find an option, a person must understand some aspects of fit (e.g. do I need something that mitigates bias in performance or outcomes? Do I need to have groups labeled?) and identify the features of their problem and prospective options to see whether they are suited. In order to evaluate fit, they must be able to predict consequences of each option; identify needed and available resources; and define (at least intuitively) success, failure, and acceptable risks.

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

1.2.2 *Judgment.*

Rest's foundational work describing what was then called "moral sensitivity" informs my conceptualization of judgment. He identifies three activities: "formulating the morally ideal course of action; deciding what one actually intends to do; or executing and implementing what one intends to do" [52].

The "moral ideal" in Rest's definition is subjective– it is ideal to the person making the judgment– but the involvement of a subjective moral ideal has implications for this study. This paper focuses on identifying ML engineers' intentions, but necessarily remains aware that some may prefer to report what they see as the moral ideal to a researcher; that in a real work situation, organizational and interpersonal factors may influence engineers' perceptions, options, and intentions; and that their perspective or options could change as they attempt to execute a decision. Therefore, although the affordances of the final design focus on presenting options and features of fit, the interface also supports two secondary goals: education about the variety of conceptions of "morally ideal" courses of action and about tools to facilitate execution in real work settings.

2 DESIGN PROBLEM

Imagine you are an ML engineer and you have recognized a potential ethical issue at work. Maybe you noticed different predictions among demographic groups in your model. You decide to learn more, so you search for a popular fairness toolkit that was recommended to you.

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

This is undoubtedly a useful resource: it offers all of the features that participants pointed to as desirable: code, tutorials, and videos. However, it may be difficult to navigate without a highly particularized understanding of the circumstance, resources, options, and consequences you are facing. If you are not familiar with what a "Manhattan Distance" is, the fact that you can determine the Manhattan distance between two distributions isn't likely to help you decide whether that is the mitigation you need. Perhaps you notice a technique that claims to improve group fairness, and you start to implement it, only to realize that it works only for groups that are explicitly defined by a feature in the data, which you can't accomplish with your data.

The goal of this project is to help machine learning engineers quickly build the understanding necessary to select an appropriate technical intervention and surface key aspects of fit to support judgment. This paper describes a Value Sensitive Design study aimed at developing a guide to ethical mitigations that considers the needs and practices of machine learning engineers while supporting the interests of lower-power stakeholders.

3 DESIGN

Value Sensitive Design uses iterative conceptual, technical, and empirical methods to develop designs that reflect the values of key stakeholders. This section describes how I used Value Sensitive Design to "design up" for higher-power actors in ML and to meet the following design goals:

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

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

These design goals were developed during conceptual investigations.

3.1 Conceptual Investigations

To identify design goals, I conducted a stakeholder analysis. The impacts of machine learning on stakeholders is a well-studied problem: I read research articles about potential problems and harms in ML [5, 54, 56], their measurement [1, 10, 13, 14, 24, 42, 45, 58, 60], values and operationalizing them [25–27, 35, 36, 61], and interventions, their motivations, and impacts [8, 9, 15, 18, 20, 28, 41, 64]. I also included some papers about generalized Artificial Intelligence, after noting that some machine learning is done to support the development of future general intelligence [34, 59]. I identified stakeholder groups and used the literature to list potential benefits and harms a guide for ML engineers could have for each, values implicated by the potential benefits and harms, and potential value conflicts among stakeholders (see: Supplementary Material). I also retained any paper that described an ethical mitigation strategy in a list so that they could be included in the ML Ethics tool.

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

Prioritizing values of lower-power stakeholders in a design for higher-power ones requires a similar "flip" to the ones used by Nader and Barabas et al. [3, 47] It prompts a switch from a defensive posture (encouraging people to protect themselves: read the privacy policy, install an extension, don't use that service) to an offensive one, in which we encourage engineers to catch and deal with potential threats. This guided the selection of ethical sensitivity as a framework and its support as a design goal. Past work in ethical sensitivity with ML engineers suggests that ML engineers may largely be ethically sensitive, may not see the recognition of ethical issues to be part of their responsibility, but respond well to signals that ethics *is* part of the task at hand. An ethical field guide, especially if used by others in their organization or occupation, can serve as such a signal, much like Datasheets seem to [12].

Holstein et al. identified several disconnects between the needs of ML engineers and the offerings of ML fairness research in 2018 [37] that helped me decide to use a series of filters. Holstein et al. identified a lack of tools about data collection (alongside a desire among engineers to intervene in data collection and curation), workers' concerns about their own blind spots about sources of unfairness, needs for proactive and holistic auditing tools, and challenges around addressing problems once they've been detected.

Before undertaking the empirical work, I conducted nine pilot interviews with ML engineers and data scientists aimed at understanding their existing training data workflows. I used this information along with another interview study focusing on the needs of ML engineers in industry [37] to form my understanding of their perspectives and practices and to guide the development of the empirical work.

3.2 Empirical Investigations

I targeted empirical investigations to collect three kinds of data: reports about how participants understood their particularization habits in the context of their occupational and organizational environment; observations of how participants particularized when first presented with ethically problematic data; and observations of participants seeking options for how to move forward. To those ends, I either asked participants direct questions about particularization in their workplace, asked them to particularize on their own (with no guide), asked them about a draft in development, or asked them to particularize using a popular AI Fairness toolkit available online. Table 1 shows how participants were distributed among these activities.

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

3.2.1 *Questions.*

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

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

3.2.2 *Particularizing without a tool.*

Eleven participants were asked to particularize on their own with the following prompt after discussing a facial recognition dataset intended to be used to identify thieves. I gave participants the following prompt: "For the next step, I'll ask you to imagine that after a few weeks of working with this data, you and your team noticed that there were a lot more men than women and that there were some skin tones not represented well in the data." I asked them to think aloud as they decided what to do next and indicated they could use the internet or any of their own resources.

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

Table 1: Empirical Investigations and Participants

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

allowed me to collect data about particularization in general among machine learning engineers.

3.2.3 Particularization with draft. The original plan for the empirical investigations was to iteratively develop a draft, giving participants prototypes as the design developed. In practice, however, I found that the low fidelity and narrowly-scoped drafts I was able to produce between sessions didn't generate meaningful data about their utility for particularization. Three participants used a draft until I determined that I could instead get particularization information by asking participants to use an existing, thorough, high-fidelity toolkit.

3.2.4 Particularization with toolkit. Six participants were asked to search for a mitigation for a fairness problem in the facial recognition data using an existing toolkit. The goal of this task was to identify barriers to search, salient or sought for features of mitigation candidates, and types of information they sought about mitigation candidates. This task collected more specific data about search practices around options.

3.3 Technical

The technical investigations integrated findings from empirical and conceptual investigations (see: Section 4 into an filterable guide to ethical mitigation strategies.

Instead of making an inexpert prototype myself, I hired a professional programmer who had experience in free and open source software, web application development, and ML. Because of his specialized knowledge in ML technology and the cultural background of engineering, he also ended up serving in a role similar to that of an informant: he offered valuable technical and usability insight to the prototype based on his experience with the ML community. Notably, he helped me phrase my filter categories more clearly. For example, I had mitigation strategies classified by "group" (including "detect," "mitigate," "plan," and "report") and ethical issues. He encouraged rewrite the filters as phrases, like "My objective is to: detect ethical issues in my model," or "mitigate an existing harm." "My ethical concern is: reducing unjust discriminatory outcomes," or "ensuring equal performance across subsets." The result of hiring this programmer is a prototype that is competently built, an interface that is legible to people in the community, and code that is comprehensible to people who would like to contribute to the open source project. [1]

4 FINDINGS

This section reviews the findings from conceptual and empirical investigations and how they were integrated in the technical investigations. For reference, screenshots of the designed artifact are included in Supplementary material, or you can interact with the tool at https://ml-ethics-tool.web.app/.

4.1 Conceptual Results

Based on the stakeholder analysis, I defined several supported values and wrote working conceptualizations for them.

Usability: the designed artifact should accomplish its other goals with minimal disruption to existing practices.

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

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

I had initially conceptualized the tool as supporting fairness in ML, but during later investigations, filters for search and exploration emerged as a way of supporting particularization while surfacing relevant mitigation strategies. Filters also allow the tool to easily scale to support multiple values in ML design without compromising usability and even include papers describing different conceptions of those values. Therefore, I expanded the scope of the tool to include privacy, accountability, and other values, and did not write working conceptualizations of supported values, preferring instead for engineers to build those conceptions as they particularize based on relevant expert opinions surfaced in their search. Another benefit of this feature is flexibility as definitions of values evolve [38] and technology and society change to generate or recognize new harms, like polarization in social media [17] and demographic representativeness in portrayals of occupations [23].

4.2 Empirical Results

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

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

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

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

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

While particularizing with an existing tool, P23 explains:

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

Participants who had less experience with facial recognition or computer vision in general used several high-level secondary sources to understand the technology. For example, P8 searched "how does facial recognition work?" while particularizing without a draft and found a video on YouTube by the same name. They scrubbed through, looking for information about how images are processed. They also searched Google for "face detection" and selected Wikipedia. After building more technical understanding, they searched Google for "bias in machine learning facial recognition," selected a Medium post, then followed a link to "Man is to Programmer as woman is to homemaker? Debiasing word embeddings." [10]

P19, P8, and P20 used or mentioned Medium.com, and P7, P8, and P20 mentioned or used towardsdatascience.com as useful sources. P20 explained their use of blogs to narrow their search for primary sources:

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

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

P8, P15, P16, and P23 mentioned looking at code. While reviewing a draft, P6 explained how time pressure informed their use of high-level sources and code:

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

I hope that by streamlining the search for mitigations, this tool can help relieve time pressure and support users thoroughly orienting to features of their context that may be important to ethical decision-making.

P4, P15, P18, and P20 used or mentioned academic papers or posters. P1, P4, P5, P6, and P10 discussed academic sources as credible and useful. Participants who discussed academic papers often engaged with questions of credibility.

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

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

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

4.2.2 *Information need.*

Participants were searching for how mitigations work and how they (don't) fit their problem and its socio-technical context.

The “How”

Regardless of what sources they sought or terms they used, when considering a solution, participants wanted to understand how the mitigation works: what they would need in order to use the mitigation strategy and what, specifically, does it *do.* The design I landed on included linking to code, papers, and “other links," which may be tutorials, videos, and demonstrations as features of the tool profile. But how could I surface the “how?" While answering questions about particularization habits, P7 offered the metaphor that inspired the final design for surfacing this essential information:

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

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

Fit

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

While looking for a mitigation without a guide, P6 discussed their use of high-level and specific sources to understand fit: "he's attached a video here, so I might just watch that and see how it works instead of having to read it all, but I usually go to just the download or just copy and paste all of this [code] right here. . .the first thing I want to do is get it running and see how it works, what it does, and what kind of changes I have to make to it to make it useful for me."

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

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

The tool I designed supports five areas of fit surfaced in the empirical investigations: objective, development stage, ML technique, data type (broadly), and ethical concern. The two areas of fit I chose not to support were application domain and detailed data types (see "Unsupported Fit" below). Filters allow searchers to find mitigations that fit all supported dimensions of fit, if they exist, or to broaden their result set by prioritizing. To further tailor fit, a tool profile supports adding additional details through tags and notes.

Development Stage Ethical mitigations in ML into three categories, based on whether they intervene in the input, the process, or the output [28]. Participants rarely included developmental stage terms (like "training data" or "before training") in their search terms, but it was frequently a part of participants' problem framing: they considered image augmentation and manipulation techniques rather than processing or post-processing mitigations.

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

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

This comment about post-processing, though, came after seven minutes of searching through the paper and discussing whether it actually intervened post-processing, or whether it *detected* problems post-processing, but actually *intervened* in training data. This motivated me to allow users to filter by intervention point "collecting/cleaning data," "training my model," and "post-training" in order highlight developmental stage as an important problem feature and help users avoid this confusion.

Objective As P18 mentioned, identifying and measuring the extent of an ethical concern can offer important information to guide the selection of a mitigation. Detection tools can also support an engineer who needs to advocate within their organization for spending some time or resources to address a problem.

These two objectives, "detect" and "mitigate," were the focus of all the options-seeking search and browsing behavior among my participants. However, two other objectives were represented in the list of interventions I compiled during the conceptual investigations, and which I thought were important to include: planning and reporting.

Papers that may help with planning include reviews of interventions in an area (e.g.,[25], papers that disambiguate important concepts and present formal models (e.g., [26]), warn of unanticipated problems (e.g., [61]), or propose a new high-level approach (e.g., [42]). Reporting papers generally offer standard documentation that can be used to describe a data set (e.g., [30]) or model (e.g [46]). Planning and reporting resources may benefit ML engineers, but engineers may not be aware of their importance, or even their existence.

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

However, selecting the options for the filter categories proved to be more complicated than I expected. Some fields, like Natural Language Processing, seemed to be fairly well-defined, both in participants' discussions and in the literature, but others were not. Should facial detection and object detection be in the same category, or separate? Responding to an interview question, P19 contended with this ambiguity:"For me personally, this is a new domain. A lot of the machine learning I have experience working with . . . they usually use biomedical or biology examples . . . still computer vision, but a different type of image."

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

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

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

Options for the data type filter were derived from the list compiled in the conceptual investigation: text data, image data, tabular data, and other data. Granular data types can be included in the tool profile.

Coding Language While particularizing with an existing tool, P18 mentioned that they prefer finding solutions that are built for the coding environment or language they work in: “But I work in MATLAB so I always use something from MATLAB.”

P19 indicated that this is a guiding feature in search: “Specifically for myself, I work a lot with Python. So [search terms] as simple as ‘Python, open source, ML’ tends to really narrow down the topics that I’m working with.”

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

Unsupported fit There are two features fit that emerged and which I did not dedicate a filter or tool profile field to: application domain (e.g., medical diagnostics) and detailed data types (e.g. MRI data). In the conceptual investigations, I encountered very few mitigations in my search that were so narrowly scoped; adding them would add many filter options or a tool profile field without much benefit. Adding even one filter (e.g., a search for “data type: X-Ray” plus “development stage: training”) would be likely to return zero results. As a compromise, the tool does include a (rudimentary, for now) search feature, which would allow a user to select the filters they want and search for application domain, granular data type, or other features of interest that may appear in the descriptive fields of the tool profile.

4.2.3 Persuasion. The stated goal of this design project is to help ML engineers understand ethical problems in their work and select appropriate mitigations. However, findings from empirical investigations harmonized with the literature, emphasizing the difficulty ML engineers face gathering support for ethical action in their organziational context [21, 43, 44, 50]. Participants discussed communication barriers that make necessary communication with clients and decision-makers difficult.

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

> “Even if engineers explain everything right, like have all the facts, know the theory, try and explain the theory and the most common like layman terms . . . If they have an idea you can’t really convince them . . . higher up people are like ‘we have a deadline to meet, we can’t do it.”’

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

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

> “This is peer reviewed. . . some of that is important. Some of it– in the business sense really isn’t: if it works, it works. I don’t really need to know all the sources necessarily. But something beyond ‘I’ve done a Kaggle[2] exercise.’ . . .if I need to tell my bosses why I spent three weeks on something that came up blank, it’s nice to not say ‘hey, a junior in high school wrote a Kaggle post on it. I thought it looked great like that.’ That is nice to have kind of some backing as to like, ‘Hey, this is the research that was going off of.’ ”

Although some participants used moral reasoning to legitimate efforts toward building unbiased algorithms, many were more comfortable with rationalization. Participants emphasized the product quality or profit impacts of a problem to argue for intervention, rather than moralization.

Referring to an anecdote about an algorithm that appeared to be predicting based on differences in where an X-Ray image came from (a city with high or low prevalence rates) rather than the important features of the image, P9 said, "Actually, I really like that one because I think it’s really instructive. It doesn’t have all the charge about like racial bias, gender bias . . . At least like we can all agree without getting into the thick of the politics, right? We don’t want false negatives in Ohio."

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

4.2.4 Evaluation. To help evaluate and improve the tool, I reached out to ML engineers and technology ethics researchers. I invited evaluators to explore the tool as they might use it in one of two roles: a user searching for "a specific potential ethical issue" or a user adding a mitigation strategy. For those testing the tool as a user seeking an ethical mitigation, I encouraged users to think of a issue they’ve encountered in their own work and offered some example ethical issues if they’d prefer, for example “imagine you have a model that will influence a consequential decision and you want to make its workings as transparent as possible to the people who will be using it” and “imagine you are in the middle of developing a model, and it doesn’t have any sensitive attributes (like race or gender) in it, but you are worried that it seems to be treating demographic groups differently anyway. You want to test and see

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

if there's something unfair going on." For those testing the tool as a user adding a mitigation strategy, asked testers to trying filling out the submission form for a strategy they are aware of that is not already represented or for one of a set of examples I offered.

Two engineers and two ethics researchers offered comments. Four suggestions were implemented based on their feedback:

- **Users should see radio buttons (instead of check boxes) when selecting filters.** Radio buttons communicate that one option can be selected at a time, where check boxes suggest that users can select more than one. Radio buttons risk implying that one option should be chosen for every filter set, which I consider to be limiting: I hope users select as few filters as are useful to maximize results returned and to encourage adapting interventions across circumstances. However, feedback made it clear that confusion generated by check boxes outweighed this concern.
- **Users should be able to see how many results are returned with each search.** "Displaying [number of results] results" at the top of the tool profiles list helps users orient themselves and understand how many options they have. It also may help people assess the landscape of options and or search for gaps.
- **Users should be able to see which filters are active when filter sets are collapsed**. The "Active filters" box was added so that users can see which filters are acting on the result set.
- **Users should be able to clear all filters with one click.** Users can now select "clear all filters."

Known Issues One of the technology ethics researchers who reviewed the tool pointed me to an article urging AI ethicists to consider the cultural and regional context when designing guidelines for AI. Before implementing this, we need to carefully consider whether the regions or countries of the publication, author(s) origin, or author(s) institutions should be considered; how to deal with multi-authored papers; and how to present and enable useful search for this information. I will seek more feedback and consider the above questions further before implementing this feature.

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

4.3 Future Development

A technology ethics researcher noted the potential for the tool to be adapted to support group work. I can imagine the interface allowing participants to bookmark and share tool profiles with one another or to collaborate on project-specific lists. I am excited about the prospect of expanding or tailoring the ML ethics tool to support group work and believe that further research is needed to understand how to support teams as they particularize and judge.

New interventions for ethical machine learning are released often. I am aware of several additional interventions that need to be added, and I am sure there are more, especially if academic papers are not published about them or if they are published in venues I am not aware of. I am eager to welcome others interested in machine learning ethics, builders of tools, and students to help expand the list of included tools. Any user can submit a tool profile through the "Contribute New Strategy" feature, where the profile draft will go to an administrator (me, for now) for approval to ensure quality.

Finally, ML engineers do not the only relatively high-power actors in this circumstance, and their actions are constrained by decisions made by their managers, product designers, clients, executives, and others. I encourage folks to consider designing up to intervene in the recognition, particularization, and judgment of other high-power actors in this system.

5 CONCLUSION

This project won't solve ML bias.

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

Fortunately, addressing bias in the design ML systems does not hinder the effort to attack bias at the source, nor will success at a better, larger justice movement render this one a waste. Just like studying up, designing up is a "yes, and" project: in order to catch and address ethical problems in a complex and dynamic social system, we need actors all along a design pipeline to be engaged. We can and should design plug-ins for users that block cookies; recruit managers, engineers, and educators who are committed to aligned technology; elect regulators who understand the technology and are interested in disrupting harms; encourage companies who invest in potentially harmful technology to protect human values; and to rigorously research the data ecosystem that fuels ML technology. We should also encourage decision-makers to consider the option not to build the system at all [4, 7].

So we won't solve bias with a field guide. What did this project accomplish? We now have a search tool for ML ethics strategies. Anyone can add to it, and using its open source code, anyone can expand, tailor, or re-purpose it. It can be used in education to raise awareness of the sources and types of ethical problems in ML: I imagine teachers assigning students to create a tool profile as a way to engage students with the technical workings of innovative ethical mitigations as well as a way to keep the tool up to date. Firms can use the tool in training, at once introducing trainees to a tool that can help them avoid ethical breeches and signaling to new hires the importance of ethical awareness in their work. It being open source will allow firms to customize it to fit their context; used this way, it could offer organizational infrastructure for ethical design [44] Researchers can use it to publicize their interventions and to identify gaps in existing ones.

This project also uncovered some features of ML Engineers' views and particularization habits that can serve researchers and designers and demonstrated that ethical sensitivity can be used as a framework to support design.

Most participants were fluent in moral evaluation as a legitimation, but sometimes hesitated to use it. Those who want to discuss ethical concerns with machine learning engineers may benefit from preferring a quality framing to discuss concerns, rather than a moral one. Another way of interpreting this finding is as call to promote moral discussions as part of engineers' education and job tasks. We found that ML engineers are often most interested in the "how" of an intervention. Marking this information clearly or surfacing it in an interface will make it easier for ML engineers to adopt an intervention. Further, if a person hoping to intervene in the development practices of ML engineers wants an ML engineer to see something, they may consider placing it somewhere in the path to "how," for example in code or tutorials

Finally, findings from this design project have implications for the product managers and executives that oversee ML development. Engineers must be be empowered to bring up potential ethical issues. Managers should work to convince engineers that they will not be replaced or punished if they express ethical concerns, but rather that their technical knowledge and ethical perceptions are valued. Give them resources about, training for, and time to implement ethical mitigations. Firms are welcome and encouraged to use the ML ethics tool designed here, or fork the project and develop one tailored to their firm or domain. ML engineers are uniquely positioned to notice, understand, and prevent potential downstream harms from the technology they build. Let them.

REFERENCES

[1] Philip Adler, Casey Falk, Sorelle A. Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing black-box models for indirect influence. 54, 1 (2018), 95–122. https://doi.org/10.1007/s10115-017-1116-3

[2] Julia Angwin and Jeff Larson. 2016. *Machine Bias.* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY, USA) *(FAT* '20)*. Association for Computing Machinery, 167–176. https://doi.org/10.1145/3351095.3372859

[4] Solon Barocas, Asia J. Biega, Benjamin Fish, Jędrzej Niklas, and Luke Stark. 2020. When not to design, build, or deploy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 695. https://doi.org/10.1145/3351095.3375691

[5] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. (2016). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899

[6] Marguerite Barry, Aphra Kerr, and Oliver Smith. 2020. Ethics on the ground: from principles to practice. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 688. https://doi.org/10.1145/3351095.3375684

[7] Eric P.S. Baumer and M. Six Silberman. 2011. When the implication is not to design (technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Vancouver BC Canada, 2271–2274. https://doi.org/10.1145/1978942.1979275

[8] Yahav Bechavod and Katrina Ligett. 2017. Learning Fair Classifiers: A Regularization-Inspired Approach. abs/1707.00044 (2017).

[9] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. (2017). arXiv:1707.00075 http://arxiv.org/abs/1707.00075

[10] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 4349–4357. http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf

[11] Karen Boyd and Katie Shilton. 2022. Adapting Ethical Sensitivity as a Construct to Study Technology Design Teams. (2022).

[12] Karen L. Boyd. 2021. Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 438:1–438:27. https://doi.org/10.1145/3479582

[13] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[14] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. 356, 6334 (2017), 183–186. https://doi.org/10.1126/science.aal4230 arXiv:1608.07187

[15] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 3992–4001. http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf

[16] Sarah Carr. 2020. 'AI gone mental': engagement and ethics in data-driven technology for mental health. 29, 2 (2020), 125–130. https://doi.org/10.1080/09638237.2020.1714011 Publisher: Routledge _eprint: https://doi.org/10.1080/09638237.2020.1714011.

[17] L. Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. 2019. Controlling Polarization in Personalization: An Algorithmic Framework. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 160–169. https://doi.org/10.1145/3287560.3287601

[18] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with Fairness Constraints. (2018). arXiv:1704.06840 http://arxiv.org/abs/1704.06840

[19] Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. 2019. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 79–88. https://doi.org/10.1145/3287560.3287587

[20] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. 2019. Slice Finder: Automated Data Slicing for Model Validation. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. 1550–1553. https://doi.org/10.1109/ICDE.2019.00139 ISSN: 2375-026X.

[21] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. 2019. Translation, Tracks & Data: an Algorithmic Bias Effort in Practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3290607.3299057

[22] Kate Crawford. 2019. Halt the use of facial-recognition technology until it is regulated. *Nature* 572, 7771 (Aug. 2019), 565–565. https://doi.org/10.1038/d41586-019-02514-7 Bandiera_abtest: a Cg_type: World View Number: 7771 Publisher: Nature Publishing Group Subject_term: Computer science, Society, Policy.

[23] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 120–128. https://doi.org/10.1145/3287560.3287572

[24] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans LA USA). ACM, 67–73. https://doi.org/10.1145/3278721.3278729

[25] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. (2017). arXiv:1702.08608 http://arxiv.org/abs/1702.08608

[26] Marina Drosou, H.v. Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. 2017. Diversity in Big Data: A Review. 5, 2 (2017), 73–84. https://doi.org/10.1089/big.2016.0054 Publisher: Mary Ann Liebert, Inc., publishers.

[27] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. (2016). arXiv:1609.07236 http://arxiv.org/abs/1609.07236

[28] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2018. A comparative study of fairness-enhancing interventions in machine learning. (2018). arXiv:1802.04422 http://arxiv.org/abs/1802.04422

[29] Batya Friedman. 1996. Value-sensitive design. 3, 6 (1996), 16–23. http://dl.acm.org/citation.cfm?id=242493

[30] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for Datasets. (2018). arXiv:1803.09010 http://arxiv.org/abs/1803.09010

[31] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on*

Fairness, Accountability, and Transparency (FAT '19)*. Association for Computing Machinery, New York, NY, USA, 90–99. https://doi.org/10.1145/3287560.3287563

[32] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. 2019. Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. https://doi.org/10.24251/HICSS.2019.258 Accepted: 2019-01-03T00:00:45Z.

[33] Hugh Gusterson. 1997. Studying Up Revisited. *PoLAR: Political and Legal Anthropology Review* 20, 1 (1997), 114–119. https://doi.org/10.1525/pol.1997.20.1.114

[34] Marc Hanheide, Moritz Göbelbecker, Graham S. Horn, Andrzej Pronobis, Kristoffer Sjöö, Alper Aydemir, Patric Jensfelt, Charles Gretton, Richard Dearden, Miroslav Janicek, Hendrik Zender, Geert-Jan Kruijff, Nick Hawes, and Jeremy L. Wyatt. 2017. Robot task planning and explanation in open and uncertain worlds. 247 (2017), 119–150. https://doi.org/10.1016/j.artint.2015.08.008

[35] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. https://openreview.net/forum?id=S1btBvZOZr

[36] Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness Without Demographics in Repeated Loss Minimization. (2018). arXiv:1806.08010 http://arxiv.org/abs/1806.08010

[37] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. 2018. Improving fairness in machine learning systems: What do industry practitioners need? (2018). https://doi.org/10.1145/3290605.3300830 arXiv:1812.05239

[38] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un)fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/3287560.3287600

[39] Lily Irani. 2019. (2019). Consortium for the Science of Sociotechnical Systems Research Summer Institute 2019.

[40] Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA) *(CHI '13)*. Association for Computing Machinery, 611–620. https://doi.org/10.1145/2470654.2470742

[41] Niki Kilbertus, Adrià Gascón, Matt J. Kusner, Michael Veale, Krishna P. Gummadi, and Adrian Weller. 2018. Blind Justice: Fairness with Encrypted Sensitive Attributes. (2018). arXiv:1806.03281 http://arxiv.org/abs/1806.03281

[42] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 656–666. http://papers.nips.cc/paper/6668-avoiding-discrimination-through-causal-reasoning.pdf

[43] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (April 2022), 52:1–52:26. https://doi.org/10.1145/3512899

[44] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376445

[45] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. 2017. Auditing Search Engines for Differential Satisfaction Across Demographics. (2017), 626–633. https://doi.org/10.1145/3041021.3054197 arXiv:1705.10689

[46] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. (2019), 220–229. https://doi.org/10.1145/3287560.3287596 arXiv:1810.03993

[47] Laura Nader. 1972. Up the Anthropologist: Perspectives Gained From Studying Up. (1972). https://eric.ed.gov/?id=ED065375

[48] Ziad Obermeyer and Sendhil Mullainathan. 2019. Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 89. https://doi.org/10.1145/3287560.3287593

[49] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown.

[50] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 7:1–7:23. https://doi.org/10.1145/3449081

[51] James Rest, Darcia Narvaez, Muriel J. Bebeau, and Stephen J. Thoma. 1999. *Postconventional moral thinking: A neo-Kohlbergian approach.* Lawrence Erlbaum Associates Publishers. Pages: ix, 229.

[52] James R. Rest. 1982. A Psychologist Looks at the Teaching of Ethics. 12, 1 (1982), 29–36. https://doi.org/10.2307/3560621

[53] Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. 2019. Integrating Ethics within Machine Learning Courses. 19, 4 (2019), 1–26. https://doi.org/10.1145/3341164

[54] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. 2015. Hidden Technical Debt in Machine Learning Systems. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 2503–2511. http://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

[55] Nick Seaver. 2014. Studying Up: The Ethnography of Technologists. http://ethnographymatters.net/blog/2014/03/10/studying-up/

[56] Andrew D. Selbst. 2017. Disparate Impact in Big Data Policing. (2017). https://doi.org/10.2139/ssrn.2819182

[57] Katie Shilton. 2013. Values Levers: Building Ethics into Design. 38, 3 (2013), 374–397. https://doi.org/10.1177/0162243912436985

[58] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. (2013). arXiv:1301.6822 http://arxiv.org/abs/1301.6822

[59] Kristinn R. Thórisson. 2007. Integrated A.I. systems. 17, 1 (2007), 11–25. https://doi.org/10.1007/s11023-007-9055-5

[60] Indrė Žliobaitė. 2015. A survey on measuring indirect discrimination in machine learning. (2015). arXiv:1511.00148 http://arxiv.org/abs/1511.00148

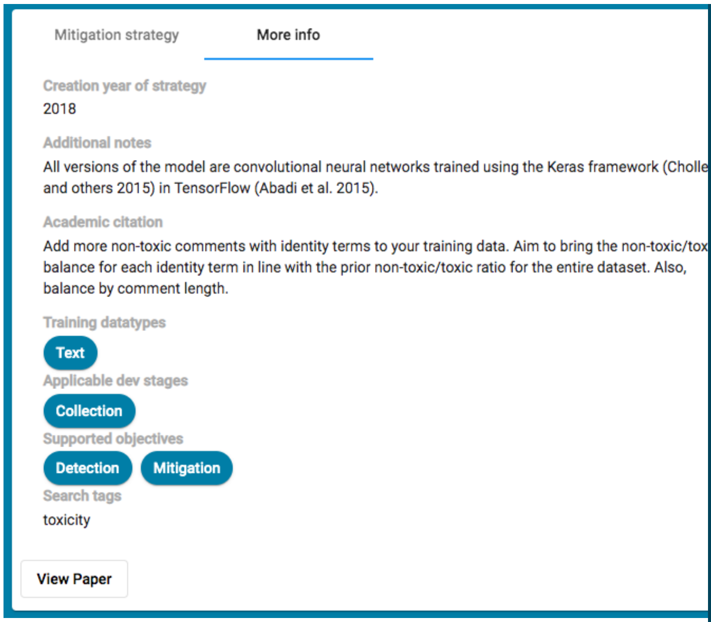
[61] Indrė Žliobaitė and Bart Custers. 2016. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. 24, 2 (2016), 183–201. https://doi.org/10.1007/s10506-016-9182-5

[62] Kathryn Weaver, Janice Morse, and Carl Mitcham. 2008. Ethical sensitivity in professional practice: concept analysis. 62, 5 (2008), 607–618. https://doi.org/10.1111/j.1365-2648.2008.04625.x

[63] Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3351095.3372833

[64] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans LA USA). ACM, 335–340. https://doi.org/10.1145/3278721.3278779

A SUPPLEMENTARY MATERIAL



Mitigation strategy | More info

Creation year of strategy
2018

Additional notes
All versions of the model are convolutional neural networks trained using the Keras framework (Cholle and others 2015) in TensorFlow (Abadi et al. 2015).

Academic citation
Add more non-toxic comments with identity terms to your training data. Aim to bring the non-toxic/tox balance for each identity term in line with the prior non-toxic/toxic ratio for the entire dataset. Also, balance by comment length.

Training datatypes
Text

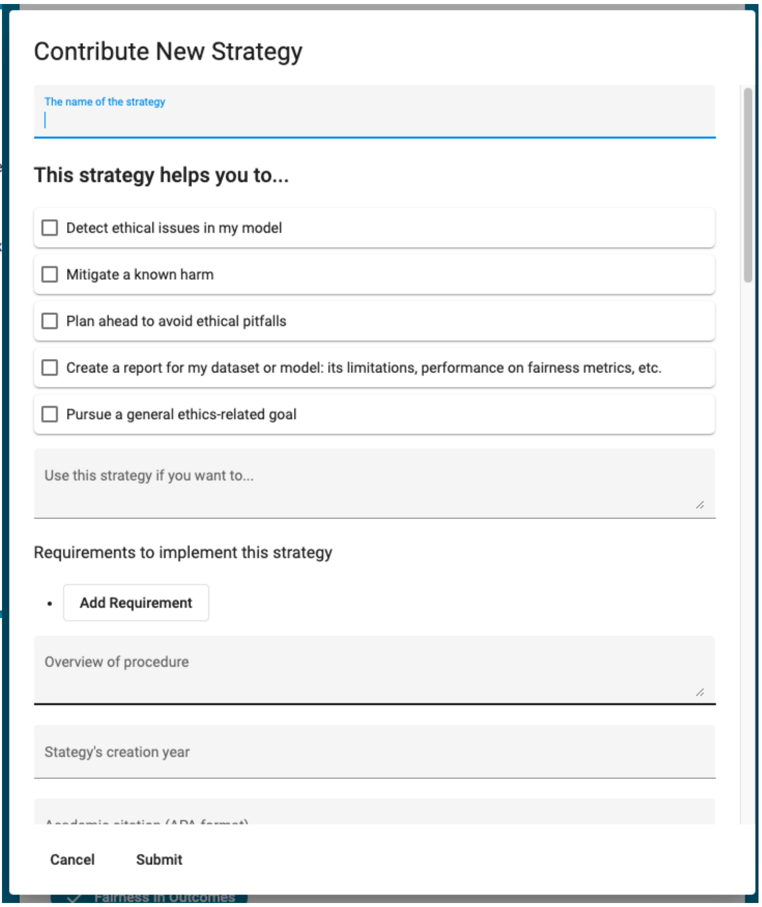
Applicable dev stages
Collection

Supported objectives
Detection | Mitigation

Search tags
toxicity

View Paper

Figure 3: Tool Profile, Additional Information



Contribute New Strategy

The name of the strategy

This strategy helps you to...

☐ Detect ethical issues in my model

☐ Mitigate a known harm

☐ Plan ahead to avoid ethical pitfalls

☐ Create a report for my dataset or model: its limitations, performance on fairness metrics, etc.

☐ Pursue a general ethics-related goal

Use this strategy if you want to...

Requirements to implement this strategy

- Add Requirement

Overview of procedure

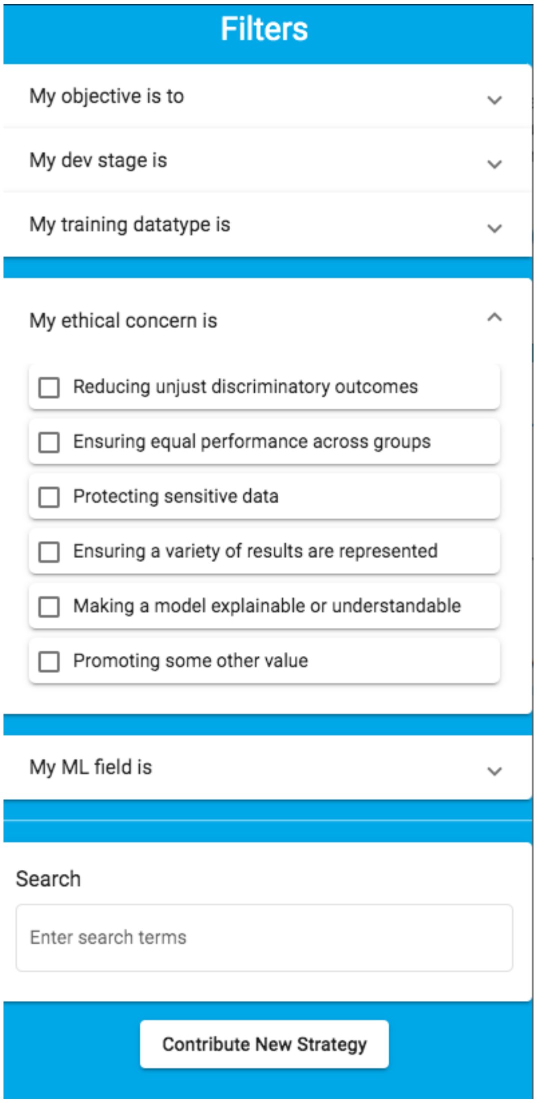
Stategy's creation year

Academic citation (APA format)

Cancel | Submit

Fairness in Outcomes

Figure 5: Adding a New Tool Profile



Filters

My objective is to

My dev stage is

My training datatype is

My ethical concern is

☐ Reducing unjust discriminatory outcomes

☐ Ensuring equal performance across groups

☐ Protecting sensitive data

☐ Ensuring a variety of results are represented

☐ Making a model explainable or understandable

☐ Promoting some other value

My ML field is

Search

Enter search terms

Contribute New Strategy

Figure 4: Filters