Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI

Michael A. Madaio
Carnegie Mellon University
Pittsburgh, PA, USA
mmadaio@cs.cmu.edu

Luke Stark
Microsoft Research
Montreal, Canada
luke.stark@microsoft.com

Jennifer Wortman Vaughan
Microsoft Research
New York, NY, USA
jenn@microsoft.com

Hanna Wallach
Microsoft Research
New York, NY, USA
wallach@microsoft.com

ABSTRACT
Many organizations have published principles intended to guide the ethical development and deployment of AI systems; however, their abstract nature makes them difficult to operationalize. Some organizations have therefore produced AI ethics checklists, as well as checklists for more specific concepts, such as fairness, as applied to AI systems. But unless checklists are grounded in practitioners' needs, they may be misused. To understand the role of checklists in AI ethics, we conducted an iterative co-design process with 48 practitioners, focusing on fairness. We co-designed an AI fairness checklist and identified desiderata and concerns for AI fairness checklists in general. We found that AI fairness checklists could provide organizational infrastructure for formalizing ad-hoc processes and empowering individual advocates. We discuss aspects of organizational culture that may impact the efficacy of such checklists, and highlight future research directions.

Author Keywords
AI; ML; ethics; fairness; co-design; checklists

CCS Concepts
•**Human-centered computing** → **Collaborative and social computing;** •**Social and professional topics** → *Codes of ethics;* •**Computing methodologies** → *Machine learning;*

INTRODUCTION
Artificial intelligence (AI) systems are increasingly ubiquitous, embedded in products and services throughout education, healthcare, finance, and beyond (e.g., [32, 69, 74]). Although these systems have enormous potential for good, they can also amplify and reify existing societal biases, such as hiring systems that are more likely to recommend applicants from certain demographic groups [4] or risk assessment systems that solidify and exacerbate patterns of racism and classism in criminal justice [6]. To mitigate such harms, many public- and private-sector organizations have published high-level principles intended to guide the ethical development and deployment of AI systems (e.g., [7, 35, 49, 56, 73]). Both concurrently and in response to these principles, researchers have created mathematical methods and software toolkits for developing fairer [8, 75, 76], more interpretable [72], and privacy-preserving AI systems [45]. Yet, amidst this activity, recent qualitative research has uncovered disconnects between the current focus of the AI ethics community and the needs of practitioners in both the public [85] and private sectors [52], contributing to broader debate around the efficacy of current approaches to AI ethics.

In particular, the abstract nature of AI ethics principles makes them difficult for practitioners to operationalize [68]. Some organizations have attempted to alleviate this difficulty by producing AI ethics checklists, as well as checklists for more specific concepts, such as fairness, as applied to AI systems (e.g., [18, 21, 27, 33, 49, 53, 60, 83, 84]). The hope is that such checklists will be an effective mechanism for ensuring that practitioners make ethical decisions as they navigate the AI development and deployment lifecycle. However, the history of checklists provides a cautionary tale. When checklists have been introduced in other domains, such as structural engineering, aviation, and medicine, without involving practitioners in their design or implementation, they have been misused or even ignored. For example, commercial airline pilots misused pre-flight checklists, resulting in catastrophe [23, 24], while surgeons initially refused to use surgical checklists [37, 47].

To understand the role of checklists in AI ethics, and specifically their role in developing and deploying fairer AI systems, we conducted an iterative co-design process with 48 practitioners from 12 technology companies, working on 37 separate products, services, or consulting engagements. Through a series of semi-structured interviews and co-design workshops, we co-designed an AI fairness checklist and identified desiderata and concerns for AI fairness checklists in general.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376445

We found that AI fairness efforts are often the result of ad-hoc processes, driven by passionate individual advocates. Organizational culture can inhibit the efficacy of these efforts. Practitioners believe that checklists could provide organizational infrastructure for formalizing ad-hoc processes and empowering individual advocates, but only if they are aligned with teams' existing workflows and supported by organizational culture. To facilitate this, we contribute desiderata and concerns for AI fairness checklists, as well as an example AI fairness checklist for practitioners to customize for their specific circumstances.

RELATED WORK

AI Ethics Principles and Software Toolkits

In recent years, a variety of public- and private-sector organizations have published principles, tenets, or value statements intended to guide the ethical development and deployment of AI systems. High-profile examples include those of the Partnership on AI [35], the European Union's High-Level Expert Group [49], and the Organisation for Economic Co-operation and Development [73]. Jobin et al. surveyed 84 distinct sets of AI ethics principles, finding that they have largely converged around a handful of more specific concepts, including fairness, accountability, transparency, and privacy, as applied to AI systems [56]. More broadly, these AI ethics principles can be situated within the history of professional and business ethics [1] and viewed as a way of shaping the "moral background" of AI [2, 88] and of computer science in general [80].

Despite their popularity, the abstract nature of AI ethics principles makes them difficult for practitioners to operationalize. Moreover, differing assumptions or interpretations can yield conflicting operationalizations—a challenge that can be masked when principles are left at a high level [68]. As a result, and in spite of even the best intentions, AI ethics principles can fail to achieve their intended goal if they are not accompanied by other mechanisms for ensuring that practitioners make ethical decisions. As Loukides et al. argue, simply having a set of principles (or even taking an oath) does not mean that those principles will be upheld in the countless decisions, both large and small, made by practitioners on a daily basis [60].

Researchers have accompanied the proliferation of AI ethics principles by creating mathematical methods and software toolkits for developing fairer [8, 75, 76], more interpretable [72], and privacy-preserving AI systems [45]. These methods and toolkits often rely on simplified, quantitative definitions of complex, nuanced concepts. For example, in the case of fairness, many researchers have focused primarily on mitigating performance disparities between racial, gender, or age groups (e.g., [3, 10, 20, 29, 71]). However, recent qualitative research focused on the needs of practitioners in the public and private sectors has uncovered significant challenges to using these methods and toolkits when developing and deploying AI systems in the real world. Veale et al. [85] found that the assumptions and interpretations of public-sector practitioners were often incompatible with quantitative definitions of fairness proposed by researchers. Similarly, Holstein et al. [52] found that private-sector practitioners felt that fairness was difficult to operationalize for many real-world AI systems, including chatbots, search engines, and automated writing evaluation systems, and could not be expressed in terms of performance disparities between groups.

Ethics is a fundamentally sociocultural concept, even when it relates to technical systems [12], meaning that AI ethics necessarily involves both sociocultural and technical factors. Pursuing purely technical solutions to AI ethics issues therefore runs the risk of committing what some researchers have referred to as a category error [77], endemic in the "technosolutionism" of computer science [66]. Such narrowly focused solutions also run the risk of participating, consciously or not, in "ethics washing:" a rhetorical commitment to addressing AI ethics issues that is unsupported by concrete actions [11]. Moreover, the ethical development and deployment of AI systems typically involves decisions that no individual practitioner can make on their own. As Stark and Hoffmann observed, AI ethics principles can place practitioners in a challenging moral bind by establishing ethical responsibilities to different stakeholders without offering any guidance on how to navigate tradeoffs when these stakeholders' needs or expectations conflict [80]. In HCI, other researchers have highlighted the role of processes and artifacts in mediating between organizations' ethics principles and UX practitioners' decisions [43]. Consistent with the literature on designing systems with human values in mind (e.g., [30, 34, 38, 39, 43, 78, 79]), these findings underscore the need for thorough, empirically grounded research on how AI ethics principles can be effectively operationalized by practitioners throughout the development and deployment lifecycle.

AI Ethics Checklists: From Principles to Practice

To help practitioners operationalize AI ethics principles, some organizations have produced AI ethics checklists, as well as checklists for more specific concepts, such as fairness, as applied to AI systems (e.g., [18, 21, 27, 33, 49, 53, 60, 83, 84]). Some of these checklists, such as the UK Department of Digital, Culture, Media and Sport's "Data Ethics Workbook" [83], were designed by public-sector organizations, while others, such as DrivenData's "Deon" ethics checklist [27], were created by private-sector technology companies; others still, such as the Johns Hopkins Center for Government Excellence's "Ethics and Algorithms Toolkit" [33], were the result of public–private partnerships. One high-profile example, created by Loukides et al. [60], drew high-level inspiration from Gawande's work on medical checklists [40].

These checklists appear to have been designed with an emphasis on breadth of applicability, so as to account for the range of contexts in which they might be implemented. Some appear to be targeted toward private-sector data scientists or AI engineers [27, 53], others toward public-sector practitioners [33, 83], while others appear to be agnostic about the intended audience [18, 49, 84]. Some checklists were designed for particular teams or organizations, and are therefore much more specific, such as Cramer et al.'s checklist for algorithmic bias in recommended content on the Spotify music streaming platform [21]. However, with the exception of Cramer et al.'s checklist, few appear to have been designed with active participation from practitioners. Different checklists are structured in different ways, with some structured by principle [53, 60, 83] and others

structured by stage of the AI development and deployment lifecycle [18, 27]. Although those structured by stage of the lifecycle may be more easily implemented by practitioners, many have obvious gaps in coverage. For example, the UK Department of Digital, Culture, Media and Sport's "Data Ethics Workbook" [83] focuses solely on data collection and analysis.

Most checklists pose items as binary "yes or no" questions, such as "Do you currently have a process to classify and assess potential risks associated with use of your product or service?" [53], thereby framing complex decisions with many competing factors as a deceptively simple compliance process. Many of these checklists appear to assume that AI ethics issues can be addressed via purely technical solutions, implemented by individual practitioners [44, 77]. This framing may be a consequence of the role of checklists in software development more broadly, where they are used to standardize known procedures for code review [15] or to serve as memory aids in guiding software security inspections [28, 31, 42]. However, AI ethics involves both sociocultural and technical factors, and concepts like fairness are often essentially contested [54]. As a result, there may not be a set of simple, agreed-upon, and known-in-advance technical actions that can be distilled into "yes or no" questions to be implemented by individual practitioners. Additionally, practitioners are often beholden to organizational constraints [65], as Gray et al. found when studying UX practitioners [43]. Therefore, the most beneficial outcome of implementing an AI ethics checklist may be to prompt discussion and reflection that might otherwise not take place [78].

Sociocultural Factors and Checklist Efficacy

In many domains, checklists are used to support task completion, guide decision making, and prompt critical conversations. In aviation and medicine, checklists often serve as memory aids [16, 47]. In structural engineering, however, checklists are used to ensure that stakeholders have met and discussed potential risks [36], many of which may not be known in advance. (Checklists have also been used as a way of promoting communication in the operating room [59].) Using checklists to prompt critical conversations is akin to treating them as "value levers"—i.e., artifacts or processes that pry open discussion about ethics [78]. As a result, the role of checklists in structural engineering may serve as a more apt inspiration for the role of checklists in AI ethics, where potential risks can be diffuse, concepts are often essentially contested, and it is not possible to guarantee that risks can be mitigated via simple technical actions. That said, there is little research on the design, implementation, and efficacy of structural engineering checklists, perhaps due to the long time spans involved or the absence of repeatable procedures, such as those inherent to aviation and medicine [36]. Indeed, research on structural engineering checklists typically limited to simulations [55].

In domains such as aviation and medicine, where checklists have been studied extensively, researchers have shown that simply having a checklist is not sufficient to influence practitioners' decisions, even for relatively simple tasks [16, 24, 87]. Although checklists were already widespread in aviation, a series of high-profile commercial airline crashes in the 1980s due to checklist misuse prompted a dedicated human-factors research program to understand barriers to their effective implementation [23]. Researchers found that the checklists themselves were only the "outer shell" of the problem: organizational processes and culture facilitated or hindered their implementation, and hence their efficacy [16, 23, 24]. For example, the checklists were often designed by airplane manufacturers, and when handed over to airline fleet or operations managers, were seldom customized for the particular organizational processes and culture of the airline [23]. Pilots and co-pilots would skip items or bundle items together when they felt that these items were redundant with other checks, thereby missing critical actions [23, 24]. The researchers proposed a set of human-centered design guidelines: establish compatibility of the checklist with organizational processes and culture; ensure consistency of the checklist (both internal consistency, as well as external consistency with other required processes and resources); develop implementation protocols (e.g., the co-pilot reads each item and the pilot confirms that it has been completed); and allow for customization after implementation [16, 24].

Despite initial success stories involving medical checklists [47, 48, 59], many studies (e.g., [13, 17, 57, 82, 87]) have highlighted the impact that sociocultural factors, including organizational processes and culture, practitioner motivation, and workflow alignment, have on the efficacy of medical checklists. For example, Borchard et al. found that aligning checklists to roles and creating opportunities for practitioner empowerment during checklist implementation were critical factors in the success of surgical checklists [13]. Weiser and Berry observed that identifying incentives for checklist use and mapping checklists to existing organization processes were also critical factors [86]. Like structural engineering checklists, they argued that the most beneficial outcomes of implementing a medical checklist may be to prompt discussion of key concerns and to empower practitioners who might otherwise not feel able to contribute to critical conversations [59, 86].

To address these challenges, some researchers have started co-designing medical checklists [14,58], drawing on participatory methods widely used in HCI (e.g., [51, 61, 91]). Borchard et al. found that checklist use and efficacy increased when stakeholders were involved in checklist design and implementation [13]. Kuo et al. co-designed a diagnostic checklist for intradialytic hypertension through a series of interviews and focus groups with nurses, clinicians, patients, and other stakeholders [58]. They elicited input on the checklist items, including tensions in the concepts covered, and identified workflow alignment. We build on this line of research, extending it to AI ethics, by co-designing an AI fairness checklist via a series of semi-structured interviews and co-design workshops.

METHODS

As discussed above, many organizations have published AI ethics principles; however, their abstract nature makes them difficult to operationalize. Some organizations have therefore produced AI ethics checklists, as well as checklists for more specific concepts, such as fairness, as applied to AI systems. Few of these checklists appear to have been designed with active participation from practitioners. Yet when checklists have

been introduced in other domains without involving practitioners in their design and implementation, they have been misused or even ignored. To understand the role of checklists in AI ethics, and specifically their role in developing and deploying fairer AI systems, we ask the following research questions:

RQ1: What are practitioners' current processes for identifying and mitigating AI fairness issues?

RQ2: What are practitioners' desiderata and concerns regarding AI fairness checklists?

RQ3: How do practitioners envision AI fairness checklists might be implemented within their organizations?

To answer these research questions, we conducted an iterative co-design process with 48 practitioners working on a variety of AI systems. We drew inspiration from Kuo et al.'s checklist co-design process [58], as well as other co-design processes, such as those used to design technologies in education [51, 61], public transit [92], and housing [91], other domains [93].

Data Collection

Semi-Structured Interviews

Our co-design process began with a series of 14 exploratory need-finding interviews, using semi-structured interview methods, to understand 1) how practitioners currently identify and mitigate AI fairness issues during their development and deployment lifecycle, and 2) what practitioners want and don't want from AI fairness checklists. Participants were shown an example surgical checklist [48], but were not shown an example AI fairness checklist. Each interview was 60–90 minutes long and was conducted via videochat. Audio was recorded and transcribed by a third-party service, in accordance with our institution's IRB; all references to specific organizations, products, services, or individuals were anonymized. We provide our interview protocol in the supplementary material.

Checklist Design

Concurrent to these interviews, two of the authors and other stakeholders in our institution designed an initial AI fairness checklist based on existing checklists and previous research, as in other checklist co-design processes (e.g., [14, 21, 58]). This initial checklist included items to consider at six different stages of the AI development and deployment lifecycle, from envisioning and defining the system to prototyping, building, launching, and evolving it. Each stage contained between six and fourteen items, such as "Envision system purpose and scrutinize for potential fairness issues," "Define and scrutinize datasets for potential fairness issues," "Define fairness criteria," and "Assess fairness criteria." The checklist was designed to guide practitioners through identifying and mitigating a variety of known AI fairness issues by soliciting input and concerns from diverse stakeholders, assessing system components for potential fairness-related harms, documenting system components, and monitoring fairness criteria, among other actions.

Co-Design Workshops

We then conducted two rounds of co-design workshops with 38 unique participants: 8 workshops, with a total of 19 participants, during the first round; and 19 workshops, with a total of 21 participants, during the second round. Each workshop was

Figure 1. Item-level checklist feedback activity.

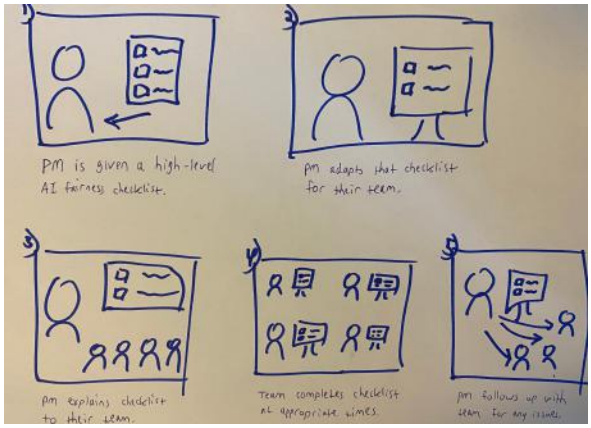

Figure 2. Example checklist implementation scenario storyboard, depicting a product manager (PM) customizing a general checklist for their team's circumstances, sharing it with their team, and ensuring its use.

90 minutes long. We used these workshops to elicit item-level feedback on our checklist, including suggestions for item revisions, additions, or removals. During the first round, we also used the checklist to elicit checklist implementation scenarios, including desiderata for and barriers to effective implementation, and then used these scenarios to generate storyboards with which to probe participants during the second round [22].

During the first round of workshops, we presented participants with all checklist items, and asked each participant to 1) place post-it notes next to items with suggestions for item revisions, additions, or removals, and 2) place colored dots on items that would be particularly easy or difficult for them (or their team or organization) to undertake. We provide an image of this activity in Figure 1. Each column represents a different stage of development and deployment lifecycle; in each column, we placed descriptions of the checklist items for that stage. We used the post-it notes and colored dots to prompt discussion around the checklist items, including items that participants felt would be particularly important or particularly difficult, as well as larger structural barriers to implementation. Following the first round, we returned to our initial checklist and revised it to reflect the participants' feedback. We also generated storyboards illustrating positive and negative checklist implementation scenarios based on the participants' comments.

During the second round of workshops, we again conducted the item-level feedback activity, this time using the revised checklist. We then showed participants the checklist implementation scenario storyboards, using them as probes in a storyboard "speed-dating" activity to rapidly elicit feedback on different scenarios [22]. During this activity, we also

Participant Roles	Participant IDs
Product or Program Manager	4, 5, 7, 8, 15, 17, 23, 26, 28, 31, 32, 36, 42, 44, 47
Data Scientist or AI/ML Engineer	9, 11, 14, 19, 29, 34, 38, 39, 40, 41, 43, 45, 48
Designer or Design Researcher	18, 20, 21, 27, 46
Software Engineer	1, 24, 30, 35, 37
Researcher	10, 16, 22, 33
Consultant	6, 12, 13
Content Editor	2, 3, 25

Table 1. Participants' roles and IDs

asked participants to generate new storyboards to illustrate their ideal implementation scenarios, as in other co-design processes [51]. We provide an example of a checklist implementation scenario storyboard in Figure 2. Following the second round, we revised our checklist again. Drawing on the semi-structured interviews and the co-design workshops, we also identified appropriate "pause points" during which the checklist items for each stage of the lifecycle might be undertaken. The final version of our checklist is provided in the supplementary material; we provide an excerpt in Figure 3.

Participants

We recruited participants primarily through snowball sampling, by posting a recruiting message on AI/ML mailing lists, Twitter, LinkedIn, and other social media platforms, and emailing contacts at technology companies. We asked people to share our recruiting message with colleagues working on AI systems. In total, we recruited 48 practitioners from 12 technology companies, working on 37 separate products, services, or consulting engagements in sectors including advertising, education, healthcare, finance, and government services. The participants span a range of technology areas, including natural language processing (10 participants), computer vision (10 participants), predictive analytics (9 participants), conversational AI or chatbots (6 participants), and information retrieval or search (6 participants). We provide the participants' roles (and IDs) in Table 1. In our post-session survey, 26 participants self-identified as men, 20 identified as women, 1 identified as non-binary, and 1 preferred not to disclose their gender to us.

Data Analysis

To analyze our interview and workshop transcripts, we adopted an inductive thematic analysis approach, modified from grounded theory for qualitative data analysis [19, 25, 70, 81]. Grounded theory is a method for emergent sense-making from data, with four stages of analysis: open coding of the raw data, generating axial codes that capture a more abstract representation of the data, organizing the axial codes into a set of categories, and summarizing the categories into "core categories" or themes [81]. Two of the authors coded the transcripts using Atlas.ti and discussed emerging themes with the other authors throughout the data collection process, collaboratively synthesizing codes as necessary to arrive at theoretical saturation—i.e., the point at which the data is fully described by the codes [81]. At each stage of our analysis, and before each of the two co-design workshop rounds, our choice of interview questions and workshop activities was shaped by our emerging understanding of the data [25, 81]. In the next three sections, we summarize our findings by describing the themes that emerged from our semi-structured interviews and co-design workshops with respect to each of our three research questions.

RQ1: PRACTITIONERS' CURRENT PROCESSES

We found that AI fairness efforts are often the result of ad-hoc processes, driven by passionate individual advocates. Organizational culture can inhibit the efficacy of these efforts by prioritizing a fast-paced development and deployment lifecycle.

Individual Advocates vs. Organizational Culture

Many participants stated they understood AI fairness through an explicitly normative lens. As one participant explained, "For me, personally, I just like doing the right thing. I think doing the right thing is the right thing, regardless of whether it makes you look bad, or whether the customer's going to come back to you or not." (P11) Alongside this framing of fairness as a personal priority, several participants acknowledged the importance of AI fairness to their organization's reputation. As one participant in a consulting organization described it,

> *One of the biggest consequences [of not addressing fairness issues] is that we're not helping our customers. I think it's our responsibility to help our customers build trust with their customers. If we don't have tools and platforms and systems that allow them to do that, we're not setting them up for success.* (P7)

Although many participants linked their organization's reputation to AI fairness, they frequently said that such efforts are largely driven by passionate individual advocates. Participants noted the challenge of finding time to adequately address AI fairness issues, given the incentives to deliver on product goals. As one participant described it, "To be honest, it felt like it was mainly up to the individuals in design and development discussions to raise awareness around fairness issues." (P17)

Individual advocates face both sociocultural barriers to speaking up and structural barriers to having their teams address AI fairness issues. Participants reported strong organizational incentives for a fast-paced development and deployment lifecycle, often standing at odds with the practice of pausing to consider fairness, similar to tensions found previously in studies of privacy [46, 62, 78]. For example, one participant told us, "I get paid to go fast. And I go as fast as I can without doing harm. We're not allowed to spend three years developing a product. We will die. Our competitors are on a weekly cadence." (P24) We also heard about conflict at an organizational level between the desire to consider AI ethics and the business imperatives of product development. "There's a broader, company-wide push-pull of 'Do I do a good thing or do I do the thing that ships the product?'" one participant noted. (P19)

Given these realities, we found social costs for individuals advocates who raise concerns about AI fairness issues and who are perceived as impeding the pace of work. Individual advocates can face criticism for posing obstacles to the perceived

Envision

Consider doing the following items in moments like:

- Envisioning meetings
- Pre-mortem screenings
- Product greenlighting meetings

1.1 Envision system and scrutinize system vision

1.1.a Envision system and its role in society, considering:

- System purpose, including key objectives and intended uses or applications
  - Consider whether the system should exist and, if so, whether the system should use AI
- Sensitive, premature, dual, or adversarial uses or applications
  - Consider whether the system will impact human rights
  - Consider whether these uses or applications should be prohibited
- Expected deployment contexts (e.g., geographic regions, time periods)
- Expected stakeholders (e.g., people who will make decisions about system adoption, people who will use the system, people who will be directly or indirectly affected by the system, society), including demographic groups (e.g., by race, gender, age, disability status, skin tone, and their intersections)
- Expected benefits for each stakeholder group, including demographic groups
- Relevant regulations, standards, guidelines, policies, etc.

1.1.b Scrutinize resulting system vision for potential fairness-related harms to stakeholder groups, considering:

- Types of harm (e.g., allocation, quality of service, stereotyping, denigration, over- or underrepresentation)
- Tradeoffs between expected benefits and potential harms for different stakeholder groups
  - Consider who the system will give power to and who it will take power from
  - Consider which expected benefits you are willing to sacrifice to mitigate potential harms

1.1.c Revise system vision to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

1.2 Solicit input and concerns on system vision

1.2.a Solicit input on system vision and potential fairness-related harms from diverse perspectives, including:

- Members of stakeholder groups, including demographic groups
  - Consider whether any stakeholder groups would prefer that the system not exist or not be deployed in all contexts, what alternatives they would prefer, and why
- Domain or subject-matter experts
- Team members and other employees

1.2.b Revise system vision to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

1.3 Escalate potential harms involving sensitive, premature, dual, or adversarial uses or applications to leadership

Figure 3. Excerpt from our AI fairness checklist: the "Envision" stage.

inexorability of the development and deployment lifecycle. As one participant described it, "There's no checkpoint where someone's supposed to say something, and so you can only do it by being this annoying squeaky wheel and, well, I'm the annoying squeaky wheel about too many things." (P19) We heard that participants wanted to advocate more strongly for AI fairness issues, but were concerned that such advocacy would have adverse impacts on their career advancement. One participant described the experience of hearing senior employees exhorting junior employees to act ethically, saying,

> *Every answer [they] gave for every question on the panel was basically, like, "Do the ethical thing and don't worry about the impact on your career." But that's an easy thing to say for a senior level person. It's a lot harder for the people in the trenches, especially when this room was full of junior designers.* (P20)

The disconnect arising from rhetorical support for AI fairness efforts coupled with a lack of organizational incentives that support such efforts is a central challenge for practitioners.

Ad-Hoc Processes

The critical role of individual advocates reflects the fact that most AI fairness efforts are the result of ad-hoc processes. As one participant noted, "At this moment it is more kind of ad hoc. If something happens, the team fixes the problem; maybe we'll fix it proactively, maybe reactively." (P31) Although many participants said that their teams do assess their systems for potential fairness-related harms, these assessments often occurred "by happenstance" (P19), and not via any kind of formal process. As one data scientist told us, "We don't have any processes or tools or anything in place to make sure that anything is fair. What happens occasionally is that one of us engineers will just spot an issue that looks like a fairness issue to us, and then we talk with each other about it, and then find some specific solution to it." (P1) Many participants echoed this sentiment, noting that AI fairness issues are often identified or mitigated by employees who happen to sit near each other or who tend to run into one another in the hallways.

RQ2: PRACTITIONERS' DESIDERATA AND CONCERNS

Participants suggested that AI fairness checklists could provide organizational infrastructure for formalizing ad-hoc processes and empowering individual advocates. However, to be most effective at achieving these goals, checklists must be aligned with teams' existing workflows, supplemented with additional resources, and not framed as a simple compliance process.

Organizational Infrastructure

Many participants said that a formal process for assessing their systems for potential fairness-related harms would help prevent issues from falling through the cracks. One participant, when shown an example of a surgical checklist [47], noted that "it would be really nice to have some sort of a checklist or something that can tell us about easy pitfalls to fall through."

(P31) Many participants described anxiety or fear that they were missing important aspects of fairness-related harms, but did not know how to assess their systems for potential risks. As one participant told us, "I think that part of the problem is people don't know what to ask. People want to do the right thing, but they don't know what the right thing is." (P27)

In addition to reducing anxiety or fear about failing to identify AI fairness issues, we found that participants believe that checklists could empower individual advocates to address such issues, or at least to raise concerns, without the social costs inherent to ad-hoc processes. Several participants analogized a checklist to a "big red button" from software security. As one participant asked, "How do we enable people to do these things without feeling like they will be labeled a troublemaker, or they will be the stop-ship person? How do we give everybody the 'big red button' without making it a problem?" (P20)

Instantiating actions for identifying and mitigating AI fairness issues in a checklist could help organizations to prioritize fairness at an organizational level. Participants described how their organizations' formal development and deployment lifecycles dictate resource allocation. As one participant said:

> *Even in a room of people who all really care, the fact that thinking about fairness isn't part of the process isn't good, because we always have more "important" priorities than we have time and resources. We have so much on our plates, and the first things to go are the ones that aren't official processes. So it doesn't matter what good intentions people have. If accounting for fairness is not a core part of the feature development process, it's not going to get done to the level of quality as things that are.* (P17)

Participants suggested that having a formal process for AI fairness efforts would allow such work to "get done to the level of quality" found elsewhere in their development and deployment lifecycle. As one participant noted, "this is an engineering company. If you establish a process, people will optimize for it and good things will happen." (P36) Again, the most beneficial outcome of implementing an AI fairness checklist may therefore be as organizational infrastructure for aligning expectations and goals around AI fairness throughout the development and deployment lifecycle. Participants saw the process of implementing a checklist as a way to spur "good tension," prompting critical conversations around AI fairness (cf. "value levers" [78]). Indeed, the checklist items that we co-designed with participants are intended to prompt these kinds of critical conversations. Like structural engineering checklists, our checklist was designed to prompt discussion and reflection that might otherwise not take place—not to remind practitioners to take simple, agreed-upon, technical actions. For example, one of our checklist item prompts teams to:

> Scrutinize system vision for potential fairness-related harms to stakeholder groups, considering: (1) types of harm (e.g., allocation, quality of service, stereotyping, denigration, over- or under-representation); and (2) trade-offs between expected benefits and potential harms for different stakeholder groups. (Checklist Item 1.1)

Workflow Alignment

When we presented participants with checklist items, we found that they felt strongly that AI fairness checklists must be aligned with teams' existing workflows. We therefore structured our checklist by stage of the AI deployment and deployment lifecycle. In contrast, many existing AI ethics checklists are structured by principle. Research on checklists in other domains, such as aviation and medicine, has indicated that identifying appropriate moments during which checklist items might be undertaken is critical for increasing checklist use [16,17,24]. We asked participants for suggestions of such moments for each stage of the lifecycle. These moments, referred to as "pause points" in the medical literature [47], differed slightly for different participants, but most participants identified moments for each stage where they were already pausing development and deployment and could potentially undertake checklist items. We provide examples of these moments in the final version of our checklist in the supplementary material.

Participants felt that some checklist items would be particularly difficult to integrate into their existing workflows, such as soliciting input and concerns from diverse stakeholders and monitoring fairness criteria after deployment. Although many participants agreed that soliciting stakeholder feedback was important, the majority believed it would be difficult to do so due to a lack of resources or even due to gaps in existing UX research methods for explicitly engaging diverse stakeholders around AI fairness. To empower teams to drive organizational change, we included checklist items prompting them to solicit stakeholder feedback at every stage of the development and deployment lifecycle, such as this item from the "Define" stage:

> Solicit input and concerns on system architecture, dataset(s), fairness criteria definitions, and potential fairness-related harms from diverse perspectives, including: (1) Members of stakeholder groups, including demographic groups; (2) Domain or subject-matter experts; (3) Team members and other employees. (Checklist Item 2.4)

When we asked participants to provided feedback on the checklist items in the final stage of the lifecycle ("Evolve"), several shared that monitoring fairness criteria after deployment would be challenging for their teams. Participants working on products or services typically said that their teams have processes for monitoring system performance or other criteria (e.g., user adoption, revenue, and speed) after deployment, but not for monitoring fairness criteria—even in cases where they already assess their systems for potential fairness-related harms before deployment. Although we included checklist items prompting teams to monitor fairness criteria after deployment, just as we added items prompting them to solicit stakeholder feedback, these findings suggest that future research is needed to develop methods and best practices in both of these areas.

Additional Resources

Many participants underscored how much they would value additional resources to supplement AI fairness checklists and help them undertake checklist items more effectively. This finding is corroborated by previous work on medical checklists, where the World Health Organization supplemented their brief

19-item surgical checklist with a larger packet of implementation guides and information [47]. Participants highlighted various items on our AI fairness checklist as places where they would find additional resources particularly useful. These resources include methods and best practices for aligning fairness criteria with the subjective experiences of users or other stakeholders and guides or templates for documenting decisions made during the development and deployment lifecycle.

For some participants, these reflections promoted analogies to other concepts, including security, accessibility, and privacy:

> *A lot of the templates that we get for filling out different parts of the [privacy review] process will have one or two examples in the table. And then the rest of the table is for us to fill in. You don't have to give me all the choices but you've got give me some framing... 'Oh, okay—it's got to look like that.' And then it says, 'Now, think of all the things...' Okay, well, I'm creative. I can think of all of the things.* (P24)

Although this participant was speaking broadly about the benefits of templates, they were doing so in reference to a checklist item that prompted teams to document dataset characteristics and limitations. The participant expressed anxiety, echoed by other participants, that documenting AI system components is particularly challenging, above and beyond the usual barriers to creating documentation during software development. Participants noted that when dealing with streaming, "eyes-off," or third-party data, documenting dataset characteristics and limitations can be especially difficult. Moreover, despite recent work on methods for documenting datasets [9, 41, 50] and models [67], similar methods do not yet exist for other components or stages of the lifecycle. Therefore, although AI fairness checklists could formalize ad-hoc processes, as described above, participants felt that they are not sufficient to do so alone and must be supplemented with additional resources.

Beyond Compliance

Although the majority of our participants reacted positively to the idea of using AI fairness checklists, we did hear concerns from several participants around the broader context of use. The primary concern was that framing AI fairness, which involves complex decisions with many competing factors, as a checklist may create the perception that it is possible to guarantee fairness by following a simple compliance process. In other words, participants saw AI fairness as a sufficiently complex, nuanced concept that a checklist could only be a starting point for efforts to engage with it. One participant said, "I'm a little bit suspicious of the checklist approach. I actually tend to think that when we have highly procedural processes we wind up with really procedural understandings of fairness." (P34) Other participants made analogies to software security checklists, where "People thought, you know, 'If I just use this security compliance checklist, I could just check things off, and then I'm good!' And they were not good." (P11) Similarly, participants shared that they were concerned that AI fairness checklists might incentivize teams to engage in the the wrong kinds of behaviors, focusing on minimal, superficial completion of items instead of engaging in discussion and reflection. Prior to engaging with our checklist, some participants were concerned that AI fairness checklists might include specific fairness criteria or thresholds to meet (our checklist does not). As one participant put it, "If any of the checklist items says, 'Have you met this number of things?' it becomes easy to game, without making things more fair." (P4) This anxiety around teams "gaming" specific criteria was tied to larger concerns regarding aspects of organizational culture that might incentivize such gaming. Another participant said, "So much of how PMs get rewarded and incentivized puts them in a position to look like they've done the right thing when maybe they're not doing the right thing." (P34) Finally, some participants were concerned that organizational culture might encourage teams to pursue purely technical solutions to issues that involve both sociocultural and technical factors. One participant told us, "That's a very non-engineering thing and the notion that engineering and technology cannot fix these problems is really upsetting to people who have spent their entire lives believing they can solve the world's problem with computing." (P34) Participants did not want AI fairness checklists to reinforce a tendency toward technosolutionism [66], and saw checklists as useful only within a broader and more holistic approach to AI fairness, relying on multiple methods and resources. As a result, our checklist items are intended to prompt critical conversations, using words like "scrutinize" and asking teams to "define fairness criteria" rather than including specific fairness criteria or thresholds to meet.

RQ3: PRACTITIONERS' IMPLEMENTATION VISIONS

Participants felt that for AI fairness checklists to be effective, they must be supported by organizational culture, customizable by teams, and integrated into organizational goals and priorities, perhaps even as metrics or key performance indicators.

Support from Organizational Culture

Although AI fairness checklists could formalize ad-hoc processes, as described above, many participants believed that such formalization would only happen if leadership changed organizational culture to make AI fairness a priority, similar to priorities and associated organizational changes made by leadership to support security, accessibility, and privacy. One participant made the analogy to internationalization, saying:

> *It's a change management problem. So I think I can only lean back on a little bit my experience with internationalization because it was very similar. When I started at our company, there was no such thing as an internationalization checklist. It was a total cowboy [situation]. People wrote code and you would try to translate the code and it would break left and right. We were breaking the company's software build with international files every day. And it took years to actually get that stuff upstream.* (P5)

According to this narrative, the practitioners writing the code to adapt to international contexts were driving organizational change upstream. This dynamic mirrors most current AI fairness efforts, where passionate individual advocates develop ad-hoc processes for their teams in the absence of top-down formal processes. However, in contrast to internationalization, these ad-hoc processes have not yet propagated upstream.

Other participants described clear messaging from leadership around AI ethics and support for disseminating best practices:

> *A CTO said, "I think you guys need to think about ethics." And that was magic, because now, all of sudden, an authority has said, "Do this thing." But I am not in a position to be enforcing this on a thousand-person org. I am in a position to write a list of questions you ought to ask yourself and publicize them as fast as I can and hope that leadership will get people to do that.* (P19)

Participants reported that the introduction of checks for accessibility issues had empowered individual advocates to pause development and deployment in order to address issues. "With accessibility, we had that checklist of things and it didn't matter if it was one person [asking to review it]. It made it very clear." (P36) This clarity minimized social costs for individual advocates who raise concerns. However, as this participant elaborated, with accessibility, teams had evolved from a state where "for a while teams wouldn't block shipping and now they definitely block shipping [if there are issues]." (P36)

Customization

Participants described how, during periods of organizational change focused on better supporting privacy or accessibility, their teams had needed to customize general policies, processes, and resources for their specific circumstances. This insight is crucial for AI fairness checklists. When we showed participants our checklist, they explained how they would need to adapt a general checklist, such as ours, to fit their team's particular needs. They also described their teams' experience adapting general accessibility and privacy checklists, with these efforts typically driven by product managers (PMs):

> *Every product team or org or division is going to have to operationalize this. And I've seen us operationalize around privacy and trust over the last five years or so because that was the last big thing we had to do. So I realize that teams do have to figure a certain amount out for themselves.* (P24)

We heard a number of suggestions for how an AI fairness checklist might support customization for the specific nuances of teams' products, services, or consulting engagements, as well as their workflows. In our storyboarding activity, we elicited feedback on different checklist implementation scenarios, many of which highlighted customization, as in the example depicted in Figure 2. One participant told us, "I think you should come up with a general fairness thinking process with guidelines for different stages, and then the PMs can follow the process and have flexibility to tailor each stage to fit what it means for the product/feature they are working on." (P29) Other participants concurred that a team lead, such as a PM or equivalent, would need to be responsible for creating a "more tailored or customized checklist." (P30) Beyond the specifics of different sectors or technology areas, participants described how AI fairness checklists would necessarily need to differ for different teams and organizations to reflect their organizational culture, goals, and priorities. As one participant put it, "the need state for fairness is ubiquitous but the right way to go about it will probably be product specific and domain specific and even organizational structure specific." (P34)

Organizational Goals and Priorities

Many participants described the importance of integrating AI fairness checklists into organizational goals and priorities. Participants in team lead roles, such as PMs or equivalent, described how their teams' performance was evaluated using a set of metrics or key performance indicators (KPIs). To justify the importance of considering fairness throughout the development and deployment lifecycle, they felt that it was critical to connect AI fairness efforts to such metrics or KPIs:

> *This is not going to be moving any of the top-line metrics that we've been used to moving for years, and not everyone may be bought in yet with the concept of this actually providing a benefit. They can see what we're doing, but it's hard to prove right now that we're helping users with this.* (P4)

This participant described how, without AI fairness efforts "moving any of the top-line metrics," they feel unable to properly justify the resources needed to address issues, given their other priorities during the development and deployment lifecycle. For this participant, there is a rhetorical need for quantifiable evidence to justify addressing AI fairness issues. Other participants echoed this sentiment, describing the importance of metrics and KPIs in negotiating priorities with leadership:

> *If I had a conversation with [our VP] and said, '[VP], would you be willing, every quarter, to see a list of all your division PMs, and have green, yellow, red on how they do on fairness, AI guidelines for customers, and things like that?' So they can say, 'Why are you red here? What's green here? How can I help you prioritize?'* (P7)

However, participants took care to note that metrics and KPIs should not reinforce a tendency toward technosolutionism.

DISCUSSION

Through an iterative co-design process, involving a series of semi-structured interviews and co-design workshops, we found that practitioners believe that AI fairness checklists could provide organizational infrastructure for formalizing ad-hoc processes. Participants suggested item-level feedback on our AI fairness checklist and voiced desiderata and concerns for AI fairness checklists in general. They highlighted that AI fairness checklists must be aligned with teams' existing workflows, supplemented with additional resources, not framed as a simple compliance process, supported by organizational culture, customizable by teams, and integrated into organizational goals and priorities. In this section, we discuss future research directions for AI fairness checklists, as well as some of the larger themes that arise from our findings.

Engaging Diverse Stakeholders

There remain several open questions and research directions that must be pursued in order to help teams introduce "good tension" [43, 78] into the AI development and deployment lifecycle so that they can engage deeply with the complex, nuanced concept of fairness, as applied to AI systems. In particular, we found that there are gaps in existing UX research

methods for explicitly engaging diverse stakeholders around AI fairness. Although participants described existing methods for user testing, current UX research methods provide little guidance on how to solicit input and concerns from stakeholders belonging to different groups, especially when some groups have substantially less power or influence than others. For example, a UX researcher working on a predictive policing system might solicit feedback from the police—i.e., the intended users of the system—but fail to engage with the communities most likely to be affected by the system's use. Even when participants reported that their teams had involved members of affected communities, they felt uncertain about how best to incorporate feedback from these communities given the power differential and the influence that paying customers typically have over decisions made during the development and deployment lifecycle. Although HCI researchers have developed methods for designing systems with human values in mind (e.g., [38,39,91,93]), they have not yet become best practices or even propagated into professional training for UX researchers. Moreover, there are significant challenges to adapting user-centered methods for AI systems (e.g., [5, 26, 89, 90]).

Different Needs

Reflective of the diversity of our participants' sectors, technology areas, and roles, we found that AI fairness checklists would need to differ for different teams or organizations to reflect their organizational culture, goals, and priorities. For example, participants from startups told us that although they would find such checklists useful for suggesting questions to ask at different stages of the development and deployment lifecycle, they had no formal process for any part of software development. In contrast, practitioners from large companies noted that they already had formal processes in place for other concepts, including security, accessibility, and privacy. We also found that current approaches to AI fairness focus primarily on products, and do little to address the needs of practitioners working on services or consulting engagements. In particular, participants raised questions about identifying and mitigating AI fairness issues relating to datasets outside of their control and about monitoring deployment contexts or fairness criteria after a system has been handed off to a customer.

Sociocultural Factors and Checklist Efficacy

We found that organizational culture typically prioritizes "moving fast" and shipping products over pausing to consider fairness, similar to tensions found previously in studies of privacy [46, 62, 78]. AI fairness checklists could therefore introduce "productive restraint" into the development and deployment lifecycle, as suggested by Matias in reference to tort law fostering innovation [63]. Participants saw the process of implementing a checklist as a way to spur "good tension," prompting critical conversations and prying open discussion about AI fairness (cf. "values levers" [78]). Mechanisms for introducing productive restraint already exist for concepts like security, accessibility, and privacy; AI fairness checklists could serve a similar role to these mechanisms, as well as providing organizational infrastructure for formalizing ad-hoc processes, thereby empowering individual advocates and minimizing social costs for raising concerns. In medicine, when surgical checklists were introduced, nurses felt more empowered to raise safety concerns before and during surgery, and were more likely to be listened to by surgeons [17, 47, 59].

Yet simply having a checklist is not sufficient to influence practitioners' decisions [43, 64, 66]. Despite acknowledging the importance of AI fairness to their organization's reputation, participants reported strong organizational incentives for a fast-paced development and deployment lifecycle. Moreover, framing AI fairness as a checklist may create the impression that it is possible to guarantee fairness by following a simple compliance process, incentivizing teams to engage in the wrong kind of behaviors or reinforcing a tendency toward technosolutionism. AI fairness checklists should therefore be supported by organizational culture and designed to prompt discussion and reflection that might otherwise not take place.

Limitations

This work is a first step toward understanding the role of checklists in AI ethics, and specifically their role in developing and deploying fairer AI systems. As an initial effort, our data is necessarily limited in both scope and coverage; we recommend that future work consider practitioners in other sectors, technology areas, and roles. We also recommend that future work consider checklists for other concepts reflected in AI ethics principles, such as accountability and transparency. Moving forward, we plan to evaluate the AI fairness checklist that we co-designed with participants by conducting pilot studies with teams, with the goal of eliciting further feedback, both on the checklist itself and on implementation scenarios.

CONCLUSION

Despite a recent proliferation of AI ethics principles, their abstract nature makes them difficult for practitioners to operationalize. Some organizations have therefore produced AI ethics checklists, as well as checklists for more specific concepts, such as fairness, as applied to AI systems. However, few appear to be designed with active participation from practitioners. We therefore conducted an iterative co-design process with 48 practitioners, in which we co-designed an AI fairness checklist and identified desiderata and concerns for AI fairness checklists in general. We found that practitioners believe that AI fairness checklists could provide organizational infrastructure for formalizing ad-hoc processes and empowering individual advocates. However, to be most effective at achieving these goals, they must be aligned with teams' existing workflows and supported by organizational culture. We identified several open questions and research directions that must be pursued in order to help teams engage deeply with the complex, nuanced concept of AI fairness, as applied to AI systems. Finally, we hope that this work inspires future efforts to co-design guided support for practitioners working to address AI ethics issues.

ACKNOWLEDGMENTS

The first author was a summer intern at Microsoft Research. We thank Saleema Amershi, Peter Bailey, Eric Charran, Natasha Crampton, Eric Horvitz, Jacquelyn Krones, Craig Shank, Steve Sweetman, and especially Mike Philips for helpful conversations, as well as other past and present members of Microsoft's Aether Working Group on Bias and Fairness.

REFERENCES

[1] Andrew Abbott. 1983. Professional Ethics. *Amer. J. Sociology* 88, 5 (March 1983), 855–885.

[2] Gabriel Abend. 2014. *The Moral Background*. Princeton University Press, Princeton, NJ and Oxford.

[3] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 2019 International Conference on Machine Learning*. 60–69.

[4] Ifeoma Ajunwa and Daniel Greene. 2019. Platforms at Work: Automated Hiring Platforms and Other New Intermediaries in the Organization of Work. In *Work and Labor in the Digital Age*. Emerald Publishing Limited, 61–91.

[5] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human–AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13. DOI:http://dx.doi.org/https://doi.org/10.1145/3290605.3300233

[6] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. 2018. Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. *Proceedings of Machine Learning Research* 81 (Jan. 2018), 1–15.

[7] Beijing Academy of Artificial Intelligence. 2019. Beijing AI Principles. (2019). https://www.baai.ac.cn/blog/beijing-ai-principles

[8] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. (Oct. 2018). https://arxiv.org/abs/1810.01943

[9] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[10] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018), 0049124118782533.

[11] Elettra Bietti. 2020. From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy. In *Proceedings of the Confernce on Fairness, Accountability, and Transparency (FAT*)*.

[12] Wiebe E. Bijker. 1997. Introduction. In *Of Bicycles, Bakelite, and Bulbs: Toward a Theory of Sociotechnical Change*. The MIT Press, Cambridge, MA, 1–17.

[13] Annegret Borchard, David LB Schwappach, Aline Barbir, and Paula Bezzola. 2012. A systematic review of the effectiveness, compliance, and critical factors for implementation of safety checklists in surgery. *Annals of surgery* 256, 6 (2012), 925–933.

[14] Paul Bowie, Julie Ferguson, Marion MacLeod, Susan Kennedy, Carl de Wet, Duncan McNab, Moya Kelly, John McKay, and Sarah Atkinson. 2015. Participatory design of a preliminary safety checklist for general practice. *Br J Gen Pract* 65, 634 (2015), e330–e343.

[15] Bill Brykczynski. 1999. A survey of software inspection checklists. *ACM SIGSOFT Software Engineering Notes* 24, 1 (1999), 82.

[16] Barbara K Burian. 2006. Design guidance for emergency and abnormal checklists in aviation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50. Sage Publications: Los Angeles, CA, 106–110.

[17] Barbara K Burian, Anna Clebone, Key Dismukes, and Keith J Ruskin. 2018. More than a tick box: medical checklist development, design, and use. *Anesthesia & Analgesia* 126, 1 (2018), 223–232.

[18] Center for Democracy and Technology. 2019. Digital Decisions Tool. (2019). https://cdt.org/blog/digital-decisions-tool/

[19] Kathy Charmaz. 2008. Grounded theory as an emergent method. *Handbook of emergent methods* 155 (2008), 172.

[20] Alexandra Chouldechova, Emily Putnam-Hornstein, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of Machine Learning Research* 81 (Jan. 2018), 1–15.

[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 *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–8. DOI:http://dx.doi.org/10.1145/3290607.3299057

[22] Scott Davidoff, Min Kyung Lee, Anind K Dey, and John Zimmerman. 2007. Rapidly exploring application design through speed dating. In *International Conference on Ubiquitous Computing*. Springer, 429–446.

[23] Asaf Degani and Earl L Wiener. 1991. Human Factors of Flight Deck Checklists: The Normal Checklist. In *NASA Contractor Report 177549; Contract NCC2-377*. National Aeronautics and Space Administration, Ames Research Center, Moffett Field, California.

[24] Asaf Degani and Earl L Wiener. 1993. Cockpit checklists: Concepts, design, and use. *Human factors* 35, 2 (1993), 345–359.

[25] Nicole M Deterding and Mary C Waters. 2018. Flexible coding of in-depth interviews: A twenty-first-century approach. *Sociological methods & research* (2018), 1–32.

[26] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 278–288. DOI: http://dx.doi.org/10.1145/3025453.3025739

[27] DrivenData. 2019. Deon: An ethics checklist for data scientists. (2019). http://deon.drivendata.org/

[28] Bob Duncan and Mark Whittington. 2014. Reflecting on whether checklists can tick the box for cloud security. In *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*. IEEE, 805–810.

[29] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. (2012), 214–226. https://arxiv.org/abs/1104.3913

[30] Cynthia Dwork and Deirdre K. Mulligan. 2013. It's Not Privacy, and It's Not Fair. *Stanford Law Review Online* 66 (Sept. 2013), 35–40.

[31] Frank Elberzhager, Alexander Klaus, and Marek Jawurek. 2009. Software inspections using guided checklists to ensure security goals. In *2009 International Conference on Availability, Reliability and Security*. IEEE, 853–858.

[32] Virgina Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, New York.

[33] Johns Hopkins Center for Government Excellence. 2019. Ethics & Algorithms Toolkit. (2019). http://ethicstoolkit.ai/

[34] Mary Flanagan and Helen Nissenbaum. 2014. *Values at Play in Digital Games*. The MIT Press, Cambridge, MA.

[35] Organisation for Economic Co-operation and Development. 2019. OECD AI Principles. (2019). https://www.oecd.org/going-digital/ai/principles/

[36] Lincoln H Forbes and Syed M Ahmed. 2010. *Modern construction: lean project delivery and integrated practices*. CRC Press, Boca Raton, FL.

[37] Aude Fourcade, Jean-Louis Blache, Catherine Grenier, Jean-Louis Bourgain, and Etienne Minvielle. 2012. Barriers to staff adoption of a surgical safety checklist. *BMJ Quality and Safety* 21, 3 (2012), 191–197.

[38] Batya Friedman and David G. Hendry. 2019. *Value Sensitive Design*. The MIT Press, Cambridge, MA.

[39] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Transactions on Information Systems* 14, 3 (Sept. 1996), 330–347.

[40] Atul Gawande. 2009. *The Checklist Manifesto: How to Get Things Right*. Metropolitan Books, New York.

[41] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for Datasets. CoRR arXiv:1803.09010. (2018).

[42] David P Gilliam, Thomas L Wolfe, Joseph S Sherif, and Matt Bishop. 2003. Software security checklist for the software life cycle. In *WET ICE 2003. Proceedings. Twelfth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, 2003.* IEEE, 243–248.

[43] Colin M Gray and Shruthi Sai Chivukula. 2019. Ethical Mediation in UX Practice. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 178. DOI: http://dx.doi.org/10.1145/3290605.3300408

[44] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. 2019. Better, Nicer, Clearer, Fairer. In *Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS)*, T X Bui and R H Sprague (Eds.). 2122–2131.

[45] Miguel Guevara. 2019. Enabling developers and organizations to use differential privacy. (September 2019). https://developers.googleblog.com/2019/09/enabling-developers-and-organizations.html

[46] Seda Gürses and Joris van Hoboken. 2017. Privacy After the Agile Turn. In *Cambridge Handbook of Consumer Privacy*, Jules Polonetsky, Omer Tene, and Evan Selinger (Eds.). CambrIdge, UK.

[47] Brigette M Hales and Peter J Pronovost. 2006. The checklist: A tool for error management and performance improvement. *Journal of Critical Care* 21, 3 (2006), 231–235.

[48] Alex B Haynes, Thomas G Weiser, William R Berry, Stuart R Lipsitz, Abdel-Hadi S Breizat, E Patchen Dellinger, Teodoro Herbosa, Sudhir Joseph, Pascience L Kibatala, Marie Carmela M Lapitan, and others. 2009. A surgical safety checklist to reduce morbidity and mortality in a global population. *New England Journal of Medicine* 360, 5 (2009), 491–499.

[49] European Union High-level Expert Group. 2019. Ethics Guidelines for Trustworthy AI: Building trust in human-centric AI. (2019). https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

[50] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. CoRR arXiv:1805.03677. (2018).

[51] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2019a. Designing for Complementarity: Teacher and Student Needs for Orchestration Support in AI-Enhanced Classrooms. In *International Conference on Artificial Intelligence in Education (AIED) 2019 Proceedings*. Springer, 157–171.

[52] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019b. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–18. DOI: http://dx.doi.org/10.1145/3290605.3300830

[53] Machine Intelligence. 2019. AI Ethics Framework. (2019). https://www.migarage.ai/ethics-framework/

[54] Abigail Z. Jacobs and Hanna Wallach. 2019. Measurement and Fairness. arXiv:1912.05511. (2019).

[55] Martin Jaeger and Desmond Adair. 2012. Communication simulation in construction management education: Evaluating learning effectiveness. *Australasian Journal of Engineering Education* 18, 1 (2012), 1–14.

[56] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* (Sept. 2019), 1–11.

[57] Heidi S Kramer and Frank A Drews. 2017. Checking the lists: A systematic review of electronic checklist use in health care. *Journal of biomedical informatics* 71 (2017), S6–S12.

[58] Pei-Yi Kuo, Rajiv Saran, Marissa Argentina, Michael Heung, Jennifer L Bragg-Gresham, Dinesh Chatoth, Brenda Gillespie, Sarah Krein, Rebecca Wingard, Kai Zheng, and Tiffany Veinot. 2019. Development of a Checklist for the Prevention of Intradialytic Hypotension in Hemodialysis Care: Design Considerations Based on Activity Theory. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14. DOI:http://dx.doi.org/10.1145/3290605.3300872

[59] Lorelei Lingard, Sherry Espin, B Rubin, Sarah Whyte, M Colmenares, GR Baker, Diane Doran, E Grober, B Orser, J Bohnen, and others. 2005. Getting teams to talk: development and pilot implementation of a checklist to promote interprofessional communication in the OR. *BMJ Quality & Safety* 14, 5 (2005), 340–346.

[60] Mike Loukides, Hilary Mason, and DJ Patil. 2018. Of Oaths and Checklists. (2018). https://www.oreilly.com/ideas/of-oaths-and-checklists

[61] Michael A Madaio, Fabrice Tanoh, Axel Blahoua Seri, Kaja Jasinska, and Amy Ogan. 2019. "Everyone Brings Their Grain of Salt": Designing for Low-Literate Parental Engagement with a Mobile Literacy Technology in Côte d'Ivoire. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–15. DOI: http://dx.doi.org/10.1145/3290605.3300695

[62] Noëmi Manders-Huits and Michael Zimmer. 2009. Values and Pragmatic Action: The Challenges of Introducing Ethical Intelligence in Technical Design Communities. *International Review of Information Ethics* (Jan. 2009), 1–8.

[63] J Nathan Matias. 2019. The American Museum of Exploding Cars and Toys That Kill You Everyone in tech should visit this museum , and so should you . (2019), 1–11. https://medium.com/berkman-klein-center/the-american-museum-of-exploding-cars-and-toys-that-kill-you-5123f35cb271

[64] Andrew McNamara, Justin Smith, and Emerson Murphy-Hill. 2018. Does ACM's code of ethics change ethical decision making in software development?. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM, 729–733.

[65] Jacob Metcalf. 2014. Ethics Codes: History, Context, and Challenges. *Council for Big Data, Ethics, and Society* (2014). http://bdes.datasociety.net/council-output/ethics-codes-history-context-and-challenges/

[66] Jacob Metcalf, Emanuel Moss, and others. 2019. Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research: An International Quarterly* 86, 2 (2019), 449–476.

[67] 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. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 220–229. DOI: http://dx.doi.org/10.1145/3287560.3287596

[68] Brent Mittelstadt. 2019. AI Ethics–Too Principled to Fail? CoRR arXiv:1906.06668, (2019).

[69] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (Sept. 2016), 205395171667967–21.

[70] Michael Muller. 2014. Curiosity, creativity, and surprise as analytic tools: Grounded theory method. In *Ways of Knowing in HCI*. Springer, 25–48.

[71] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. New York.

[72] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223* (2019).

[73] Partnership on AI. 2017. AI Tenets. *Partnership on AI* (2017). https://www.partnershiponai.org/2016/09/industry-leaders-establish-partnership-on-ai-best-practices/

[74] Cathy O'Neil. 2017. *Weapons of Math Destruction*. Broadway Books, New York.

[75] Google Research. 2019. What if Tool. (2019). https://pair-code.github.io/what-if-tool/

[76] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. 2018. Aequitas: A Bias and Fairness Audit Toolkit. 2018 (2018). http://arxiv.org/abs/1811.05577

[77] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. (2019), 59–68.

[78] Katie Shilton. 2013. Values levers: Building ethics into design. *Science, Technology, & Human Values* 38, 3 (2013), 374–397.

[79] Katie Shilton. 2018. Values and Ethics in Human-Computer Interaction. *Foundations and Trends in Human–Computer Interaction* 12, 2 (2018), 107–171.

[80] Luke Stark and Anna Lauren Hoffmann. 2019. Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture. *Journal of Cultural Analytics* (May 2019).

[81] Anselm Strauss and Juliet M Corbin. 1990. *Basics of qualitative research: Grounded theory procedures and techniques.* Sage Publications, Inc.

[82] Øyvind Thomassen, Ansgar Storesund, Eirik Søfteland, and Guttorm Brattebø. 2014. The effects of safety checklists in medicine: a systematic review. *Acta Anaesthesiologica Scandinavica* 58, 1 (2014), 5–18.

[83] United Kingdom, Department of Digital, Culture, Media and Sport. 2019. Data Ethics Workbook. (2019). https://www.gov.uk/government/publications/data-ethics-workbook/data-ethics-workbook

[84] Shannon Vallor. 2019. An Ethical Toolkit for Engineering / Design Practice. (2019). https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/

[85] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14. DOI:http://dx.doi.org/10.1145/3173574.3174014

[86] Thomas G Weiser and William R Berry. 2013. Perioperative checklist methodologies. *Canadian Journal of Anesthesia/Journal canadien d'anesthésie* 60, 2 (2013), 136–142.

[87] Thomas G Weiser, Alex B Haynes, Angela Lashoher, Gerald Dziekan, Daniel J Boorman, William R Berry, and Atul A Gawande. 2010. Perspectives in quality: designing the WHO Surgical Safety Checklist. *International journal for quality in health care* 22, 5 (2010), 365–370.

[88] Christine T Wolf. 2019. Conceptualizing Care in the Everyday Work Practices of Machine Learning Developers. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*. ACM, 331–335. DOI: http://dx.doi.org/10.1145/3301019.3323879

[89] Qian Yang. 2018. Machine Learning as a UX Design Material: How Can We Imagine Beyond Automation, Recommenders, and Reminders?. In *2018 AAAI Spring Symposium Series*.

[90] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating how experienced UX designers effectively work with machine learning. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, 585–596. DOI: http://dx.doi.org/10.1145/3196709.3196730

[91] Daisy Yoo, Alina Huldtgren, Jill Palzkill Woelfer, David G Hendry, and Batya Friedman. 2013. A value sensitive action-reflection model: evolving a co-design space with stakeholder and designer prompts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 419–428. DOI: http://dx.doi.org/10.1145/2470654.2470715

[92] Daisy Yoo, John Zimmerman, Aaron Steinfeld, and Anthony Tomasic. 2010. Understanding the space for co-design in riders' interactions with a transit service. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1797–1806. DOI: http://dx.doi.org/10.1145/1753326.1753596

[93] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-sensitive algorithm design: method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction* 2 (2018), 194. DOI:http://dx.doi.org/10.1145/3274463

Preamble

Fairness is a complex concept and deeply contextual. Keep the following points in mind:

- There is no single definition of fairness that will apply equally well to different applications of AI.
- Given the many complex sources of unfairness, it is not possible to fully "debias" a system or to guarantee fairness; the goal is to detect and to mitigate fairness-related harms as much as possible.
- Prioritizing fairness in AI systems often means making tradeoffs based on competing priorities. It is therefore important to be explicit and transparent about priorities and assumptions.
- There are seldom clear-cut answers. It is therefore important to document your processes and considerations (including priorities and tradeoffs), and to seek help when needed.
- Detecting and mitigating fairness-related harms requires continual attention and refinement.
- If you do not feel you can detect or mitigate fairness-related harms sufficiently, seek help.

Prioritizing fairness in AI systems is a *sociotechnical* challenge. AI systems can behave unfairly for a variety of reasons, some social, some technical, and some a combination of both social and technical.

- AI systems can behave unfairly because of societal biases reflected in the datasets used to trained them.
- AI systems can behave unfairly because of societal biases that are either explicitly or implicitly reflected in the decisions made by teams during the AI development and deployment lifecycle.
- AI systems can possess characteristics that, while not necessarily reflective of societal biases, can still result in unfair behavior when these systems interact with particular stakeholders after deployment.

AI systems can cause a variety of fairness-related harms, including harms involving people's individual experiences with AI systems or the ways that AI systems represent the groups to which they belong.

- AI systems can unfairly allocate opportunities, resources, or information.
- AI systems can fail to provide the same quality of service to some people as they do to others.
- AI systems can reinforce existing societal stereotypes.
- AI systems can denigrate people by being actively derogatory or offensive.
- AI systems can over- or underrepresent groups of people, or even treat them as if they don't exist.

These types of harm are not mutually exclusive; a single AI system can exhibit more than one type.

Fairness-related harms can have varying severities. However, the cumulative impact of even comparatively "non-severe" harms can be extremely burdensome or make people feel singled out or undervalued.

Identifying who is at risk of experiencing fairness-related harms involves considering both the people who will use the system and the people who will be directly or indirectly affected by the system, either by choice or not. Although fairness is often discussed with respect to groups of people who are protected by anti-discrimination laws, such as groups defined in terms of race, gender, age, or disability status, the most relevant groups are often context-specific. Moreover, such groups may be difficult to identify. It can therefore be useful to consider the system's purpose and expected deployment contexts; different stakeholders, including the people who are responsible for, will use, or will be affected by the system, as well as the different demographic groups represented by these stakeholders; and any relevant standards, regulations, guidelines, or policies. Finally, people often belong to overlapping groups—different combinations of race, gender, and age, for example—and specific intersectional groups may be at greatest risk of experiencing fairness-related harms and at risk of experiencing different types of harm. Considering each group separately from the others may obscure these harms.

For more information about this checklist, please see M. Madaio, L. Stark, J. W. Vaughan, and H. 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 (CHI 2020).

AI Fairness Checklist

The items in this checklist are intended to be used as a starting point for teams to customize. Not all items will be applicable to all AI systems, and teams will likely need to add, revise, or remove, items to better fit their specific circumstances. Undertaking the items in this checklist will not guarantee fairness. The items are intended to prompt discussion and reflection. Most items can be undertaken in multiple different ways and to varying degrees.

Envision

Consider doing the following items in moments like:

- Envisioning meetings
- Pre-mortem screenings
- Product greenlighting meetings

1.1 Envision system and scrutinize system vision

1.1.a Envision system and its role in society, considering:

- System purpose, including key objectives and intended uses or applications
  - Consider whether the system should exist and, if so, whether the system should use AI
- Sensitive, premature, dual, or adversarial uses or applications
  - Consider whether the system will impact human rights
  - Consider whether these uses or applications should be prohibited
- Expected deployment contexts (e.g., geographic regions, time periods)
- Expected stakeholders (e.g., people who will make decisions about system adoption, people who will use the system, people who will be directly or indirectly affected by the system, society), including demographic groups (e.g., by race, gender, age, disability status, skin tone, and their intersections)
- Expected benefits for each stakeholder group, including demographic groups
- Relevant regulations, standards, guidelines, policies, etc.

1.1.b Scrutinize resulting system vision for potential fairness-related harms to stakeholder groups, considering:

- Types of harm (e.g., allocation, quality of service, stereotyping, denigration, over- or underrepresentation)
- Tradeoffs between expected benefits and potential harms for different stakeholder groups
  - Consider who the system will give power to and who it will take power from
  - Consider which expected benefits you are willing to sacrifice to mitigate potential harms

1.1.c Revise system vision to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

1.2 Solicit input and concerns on system vision

1.2.a Solicit input on system vision and potential fairness-related harms from diverse perspectives, including:

- Members of stakeholder groups, including demographic groups
  - Consider whether any stakeholder groups would prefer that the system not exist or not be deployed in all contexts, what alternatives they would prefer, and why
- Domain or subject-matter experts
- Team members and other employees

1.2.b Revise system vision to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

1.3 Escalate potential harms involving sensitive, premature, dual, or adversarial uses or applications to leadership

Define

Consider doing the following items in moments like:

- Spec reviews
- Game plan reviews
- Design reviews

For more information about this checklist, please see M. Madaio, L. Stark, J. W. Vaughan, and H. 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 (CHI 2020).

2.1 Define and scrutinize system architecture

2.1.a Define system architecture, considering:

- Machine learning models, including their structures, relationships, and interactions
- Objective functions and training algorithms
- Performance metrics (e.g., accuracy, user satisfaction, relevance)
- Functionality for stakeholder feedback (e.g., comments or concerns, third-party audits)
- Functionality for rollback or shutdown in the event of unanticipated fairness-related harms
- Functionality for preventing any prohibited uses or applications
- User interfaces or user experiences
- Other hardware, software, or infrastructure
- Assumptions made when operationalizing system vision via system architecture
    - Consider whether these assumptions are sufficiently well justified

2.1.b Scrutinize resulting definitions for potential fairness-related harms to stakeholder groups, considering:

- Types of harm (e.g., allocation, quality of service, stereotyping, denigration, over- or underrepresentation)
- Tradeoffs between expected benefits and potential harms for different stakeholder groups

2.1.c Revise system architecture definitions to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

2.2 Define and scrutinize datasets

2.2.a Define datasets needed to develop and test the system, considering:

- Desired quantities and characteristics, considering:
    - Relevant stakeholder groups, including demographic groups
        - Consider oversampling smaller stakeholder groups, but be aware of overburdening
    - Expected deployment contexts
- Potential sources of data
    - Consider reviewing all datasets from third-party vendors
- Collection, aggregation, or curation processes, including:
    - Procedures for obtaining meaningful consent from data subjects
    - People involved in collection, aggregation, or curation, including demographic groups
        - Consider whether people involved might introduce societal biases
    - Incentives for data subjects and people involved in collection, aggregation, or curation
        - Consider whether data subjects might feel undue pressure to provide data
    - Software, hardware, or infrastructure involved in collection, aggregation, or curation
- Relevant regulations, standards, guidelines, policies, etc.
- Assumptions made when operationalizing system vision via datasets
    - Consider whether these assumptions are sufficiently well justified

Scrutinize resulting definitions for potential fairness-related harms to stakeholder groups, considering:
Types of harm (e.g., allocation, quality of service, stereotyping, denigration, over- or underrepresentation)

- Tradeoffs between expected benefits and potential harms for different groups

2.2.b Revise dataset definitions to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

2.3 Define and scrutinize fairness criteria

2.3.a Based on potential fairness-related harms identified so far, define fairness criteria, considering:

- How criteria will be assessed (e.g., fairness metrics and benchmark dataset, system walkthroughs with diverse stakeholders or personas) at each subsequent stage of the lifecycle, including
    - People involved in assessment (e.g., judges), including demographic groups
        - Consider whether people involved might introduce societal biases
    - Datasets needed to assess fairness criteria
- Acceptable (levels of) deviation from fairness criteria
- Potential adversarial threats or attacks to fairness criteria (e.g., “brigading”)
- Assumptions made when operationalizing system vision via fairness criteria

For more information about this checklist, please see M. Madaio, L. Stark, J. W. Vaughan, and H. 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 (CHI 2020).

    - Consider whether these assumptions are sufficiently well justified

2.3.b Scrutinize fairness criteria definitions for potential fairness-related harms that may not be covered

2.3.c Revise fairness criteria definitions to cover any not-covered potential harms; if this is not possible, document why, along with contingency plans, etc., and consider aborting development

2.4 Solicit input and concerns on system architecture, dataset, and fairness criteria definitions

2.4.a Solicit input on definitions and potential fairness-related harms from diverse perspectives, including:

- Members of stakeholder groups, including demographic groups
- Domain or subject-matter experts
- Team members and other employees

2.4.b Revise definitions to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

Prototype

Consider doing the following items in moments like:

- Go / no-go discussions
- Code reviews

3.1 Prototype (and scrutinize) datasets

3.1.a Prototype datasets according to dataset definitions; if datasets deviate from definitions during development, revisit checklist items from "define" stage

3.1.b Document dataset characteristics and limitations (e.g., by creating datasheets), considering:

- Potential audiences for documentation, including:
    - Members of stakeholder groups
    - Team members and other employees
    - Regulators and other third parties

3.2 Prototype (and scrutinize) system

3.2.a Prototype system according to system architecture definitions; if system architecture deviates from definitions during development, revisit checklist items from "define" stage

3.2.b Document system characteristics and limitations (e.g., by creating model cards for the models that comprise the system or a transparency note or factsheet for the system itself), considering:

- Potential audiences for documentation, including:
    - Members of stakeholder groups
    - Team members and other employees
    - Regulators and other third parties

3.3 Assess fairness criteria

3.3.a Assess fairness criteria according to fairness criteria definitions, considering:

- Acceptable (levels of) deviation from fairness criteria
- Tradeoffs between different fairness criteria
- Tradeoffs between performance metrics and fairness criteria
- Discrepancies between development environment and expected deployment contexts

3.3.b If system prototype fails to satisfy fairness criteria, revise system accordingly; if this is not possible, document why, along with contingency plans, etc., and consider aborting development

3.4 Undertake user testing

3.4.a Undertake user testing with diverse stakeholders, analyzing results broken down by relevant stakeholder groups. This should be done even if the system satisfies the fairness criteria because the system may exhibit unanticipated fairness-related harms not covered by the fairness criteria. Consider conducting:

- Online experiments
- Ring testing or dogfooding
- Field trials or pilots in deployment contexts

3.4.b Revise production system to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

For more information about this checklist, please see M. Madaio, L. Stark, J. W. Vaughan, and H. 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 (CHI 2020).

3.5 Solicit input and concerns on system prototype

3.5.a Solicit input on system prototype from diverse perspectives, including:

- Members of stakeholder groups, including demographic groups
- Domain or subject-matter experts
- Team members and other employees

3.5.b Revise system prototype to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

Build

Consider doing the following items in moments like:

- Go / no-go discussions
- Code reviews
- Ship reviews
- Ship rooms

4.1 Build (and scrutinize) production datasets

4.1.1 Build production datasets according to dataset definitions; if datasets deviate from definitions during development, revisit checklist items from "define" stage

4.1.2 Update dataset documentation

4.2 Build (and scrutinize) production system

4.2.1 Build production system according to system architecture definitions; if system architecture deviates from definitions during development, revisit checklist items from "define" stage

4.2.2 Update system documentation

4.3 Assess fairness criteria

4.3.1 Assess fairness criteria according to fairness criteria definitions, considering

- Acceptable (levels of) deviation from fairness criteria
- Tradeoffs between different fairness criteria
- Tradeoffs between performance metrics and fairness criteria
- Discrepancies between development environment and expected deployment contexts

4.3.2 If production system fails to satisfy fairness criteria, revise system accordingly; if this is not possible, document why, along with contingency plans, etc., and consider aborting development

4.4 Undertake user testing

4.4.1 Undertake user testing with diverse stakeholders, analyzing results broken down by relevant stakeholder groups. This should be done even if the system satisfies the fairness criteria because the system may exhibit unanticipated fairness-related harms not covered by the fairness criteria. Consider conducting:

- Online experiments
- Ring testing or dogfooding
- Field trials or pilots in deployment contexts

4.4.2 Revise production system to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

4.5 Solicit input and concerns on production system

4.5.1 Solicit input on production system from diverse perspectives, including:

- Members of stakeholder groups, including demographic groups
- Domain or subject-matter experts
- Team members and other employees

4.5.2 Revise production system to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development

For more information about this checklist, please see M. Madaio, L. Stark, J. W. Vaughan, and H. 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 (CHI 2020).

Launch

Consider doing the following items in moments like:

- Ship review before launch
- Code reviews

5.1 Participate in public benchmarks

5.1.a Participate in public benchmarks so that stakeholders can contextualize system performance, considering:

- Competitors' responsible AI principles and development practices
- Alternatives to public benchmarks if relevant public benchmarks don't exist (e.g., distributing and publicizing private benchmark datasets for use by competitors or third parties)

5.1.b Revise system to mitigate any harms revealed by benchmarks; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting deployment

5.2 Enable functionality for stakeholder feedback

5.2.a Establish processes for responding to or escalating stakeholder feedback, including:

- Stakeholder comments or concerns
  - Consider establishing processes for redress
- Third-party audits

5.3 Enable functionality for rollback or shutdown in the event of unanticipated fairness-related harms

5.3.a Establish processes for deciding when to roll back or shut down

5.4 Enable functionality to prevent prohibited uses or applications

5.4.a Establish processes for deciding whether unanticipated uses or applications should be prohibited

Evolve

Consider doing the following items in moments like:

- Regular product review meetings
- Code reviews

6.1 Monitor deployment contexts

6.1.a Monitor deployment contexts for deviation from expectations, including:

- Unanticipated stakeholder groups, including demographic groups
- Adversarial threats or attacks

6.1.b Revise system (including datasets) to match actual deployment contexts; if this is not possible, document why, along with expected impacts on stakeholders, and consider rollback or shutdown

6.2 Monitor fairness criteria

6.2.a Monitor fairness criteria for deviation from expectations, including:

- Adversarial threats or attacks

6.2.b If system fails to satisfy fairness criteria, revise system accordingly; if this is not possible, document why, along with expected impacts on stakeholders, and consider rollback or shutdown

6.3 Monitor stakeholder feedback

6.3.a Follow processes for responding to or escalating stakeholder feedback

6.3.b Revise system to mitigate any harms revealed by stakeholder feedback; if this is not possible, document why, update system documentation, and consider rollback or shutdown

6.4 Revise system at regular intervals to capture changes in societal norms and expectations

6.4.a Revisit checklist items from previous stages

For more information about this checklist, please see M. Madaio, L. Stark, J. W. Vaughan, and H. 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 (CHI 2020).