Prompting Conversations about Fairness in AI Development with Checklists

Michael Madaio
michael.madaio@microsoft.com
Microsoft Research
New York, NY

Luke Stark
cstark23@uwo.ca
University of Western Ontario
London, Ontario

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

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

ABSTRACT

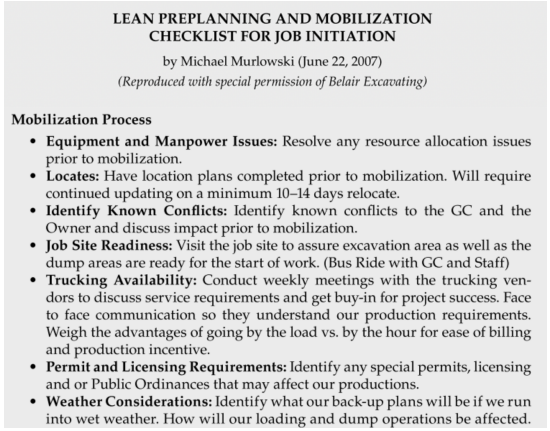
In many high-stakes domains, such as healthcare and aviation, checklists have had significant, positive impacts. However, checklists may be inappropriate when used as compliance-oriented memory aids in domains where the actions and decisions of stakeholders are contingent on and situated within particular sociocultural contexts, such as the ethical development and deployment of AI systems. In this paper, we therefore draw inspiration from the role that checklists play in the construction industry, where they are used to make sure that stakeholders communicate with one another. We study how checklists can serve as "values levers" in enabling and structuring conversations about ethics, as well as more specific concepts, such as fairness, when developing and deploying AI systems. To do this, we use data from 48 practitioners who were previously involved in co-designing an AI fairness checklist.

KEYWORDS

AI, ML, ethics, fairness, checklists

INTRODUCTION

AI systems are increasingly ubiquitous, even in high-stakes domains such as education, employment, and healthcare. Although AI systems have the potential for positive impacts, they can also reproduce or amplify societal inequities and cause other harms [25]. To mitigate these (and other) risks, numerous



LEAN PREPLANNING AND MOBILIZATION CHECKLIST FOR JOB INITIATION

by Michael Murlowski (June 22, 2007)

(Reproduced with special permission of Belair Excavating)

Mobilization Process

- **Equipment and Manpower Issues:** Resolve any resource allocation issues prior to mobilization.
- **Locates:** Have location plans completed prior to mobilization. Will require continued updating on a minimum 10–14 days relocate.
- **Identify Known Conflicts:** Identify known conflicts to the GC and the Owner and discuss impact prior to mobilization.
- **Job Site Readiness:** Visit the job site to assure excavation area as well as the dump areas are ready for the start of work. (Bus Ride with GC and Staff)
- **Trucking Availability:** Conduct weekly meetings with the trucking vendors to discuss service requirements and get buy-in for project success. Face to face communication so they understand our production requirements. Weigh the advantages of going by the load vs. by the hour for ease of billing and production incentive.
- **Permit and Licensing Requirements:** Identify any special permits, licensing and or Public Ordinances that may affect our productions.
- **Weather Considerations:** Identify what our back-up plans will be if we run into wet weather. How will our loading and dump operations be affected.

Figure 1: Construction checklist excerpted from Forbes and Ahmed [10].

"*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)

private- and public-sector organizations have published principles intended to guide the ethical development and deployment of AI systems (see Jobin et al. for a review [19]). However, even with these principles in place, AI practitioners report that fairness issues affecting their systems are often only discovered after deployment [17]. As a result, some organizations have developed checklists (e.g., [22]) intended to help teams anticipate and address ethical issues, including fairness issues, during the development and deployment process. These checklists (e.g., [3, 7, 9, 16, 18, 22, 34]) appear to draw inspiration from checklists in software development (e.g., [4, 8, 14, 35]; see [2] for a review) and in other domains, such as healthcare (e.g., [15, 20]) and aviation (e.g., [5, 6]). Many of them seem to have been designed to serve as compliance-oriented memory aids, in which the ethical development and deployment of AI systems is framed as a sequence of discrete technical actions to be completed by individual practitioners.

In practice, however, AI systems are developed and deployed via a collaborative, negotiated process, involving multiple stakeholders on the same and different teams [26, 27]. Moreover, the decisions that practitioners make throughout this process, which shape the resulting systems, are contingent on and situated within particular sociocultural contexts. These contexts give rise to numerous "micro-ethical" decisions that cannot be specified in advance, yet whose cumulative impacts can be significant [33]. Given the collaborative, negotiated, and contextual nature of the AI development and deployment process, a more relevant source of inspiration might be the checklists found in domains where stakeholders' perspectives are situated within broader contexts, such as the construction industry. In these domains, checklists do not serve as compliance-oriented memory aids, but instead prompt and guide critical conversations between stakeholders [10, 13, 28, 32]. As Gawande puts it, "[the checklist] didn't specify construction tasks; it specified communication tasks" [13]. Indeed, even in domains where checklists are primarily used as memory aids, they have also been shown to be effective in "getting teams to talk" [21].

To better understand the role that checklists might play in the ethical development and deployment of AI systems, we previously co-designed an AI fairness checklist with 48 practitioners from 12 technology companies, using a co-design process to identify desiderata and concerns for AI fairness checklists in general [23]. Via an inductive thematic analysis approach, we found that practitioners felt that organizational culture, including incentives, inhibited them from advocating for and addressing fairness issues, and that checklists could provide organizational infrastructure for supporting their efforts (see quote from P17 in the sidebar). However, we also found that practitioners were concerned that checklists might lead to compliance-oriented, procedural thinking about fairness, potentially even incentivizing teams to prioritize superficial technical actions over deeper engagement. In this paper, we revisit the data from our co-design process to specifically identify themes around the ways that AI fairness checklists might prompt and guide critical conversations between stakeholders (*cf.* [13, 21, 28]).

FINDINGS

Instead of serving as compliance-oriented memory aids, intended to facilitate a sequence of discrete technical actions to be completed by individual practitioners, checklists *could* prompt and guide critical conversations between stakeholders about fairness and other concepts relating to ethics. In particular, we found that practitioners felt that fairness checklists could empower them to raise concerns about fairness issues that they might not otherwise feel comfortable raising. Many of them stated that their teams' existing development and deployment process did not include anticipating and addressing fairness issues, and that a fairness checklist (or some other formal process relating to fairness) would help them to focus on fairness issues both in abstract discussions at the start of a project and in concrete discussions about specific system components, including datasets, later on. As one participant put it, a checklist could "*force the articulation of many of the things embedded in here* [points to checklist]. *But we don't do that as a matter of course*" (P12). Others echoed the value of a checklist in "*making people stop to think*" (P25), saying that teams "*should be forced to have that conversation*" (P5). In this way, fairness checklists can serve as "values levers" [30, 31] in enabling and structuring conversations about fairness throughout the development and deployment process.

"*On our team this* [points at Prototyping phase of fairness checklist] *is all done by the engineering team. We might have some input but, I think we're always kind of fighting for a seat at the table to be honest. We have engineering that works kind of separately, so we try to do this and I'm sure that they try to do this as well, but I'm just saying I think if we were to work together a little bit more closely, I think we'd be much better off.*" (P14)

We found that although practitioners with more experience in areas adjacent to value-sensitive design, such as user researchers, may be best able to support work relating to fairness issues, they are not typically empowered to raise these issues or to ensure that other team members work to address them. Meanwhile, although the data scientists and ML engineers with whom they work are "*aware that they need to be thinking about*" fairness (P25), many feel it is not their responsibility or beyond their expertise, perhaps as a result of their training, professional identity, or organizational role (*cf.* [24, 29]). Indicative of this reticence, one participant suggested that fairness checklists could "*make it less taboo within engineering teams*" to talk about fairness (P7). As illustrated by the quote in the sidebar, another participant, a user researcher, noted that their teams' efforts were not as effective as they could be because the user researchers were "*fighting for a seat at the table*" (P14). Other participants shared that even on teams with closer connections between engineering, user research, and business, team members do not collaborate on anticipating and addressing fairness issues. Nonetheless, we found that practitioners wanted this work to be collaborative, stressing the importance of collaboration in ensuring mutual accountability and buy-in. Practitioners also noted the need for diverse perspectives and skillsets when engaging with sociotechnical concepts like fairness. In their work on the collaborative, negotiated, and contextual nature of the development and deployment of AI systems, Passi and Sengers call for orienting towards the organizational dynamics that shape those collaboration processes [27].

Rather than laying out a sequence of discrete technical actions known to address specific fairness issues, checklists could be used to prompt teams to come together at crucial points during the development and deployment process to collaboratively anticipate and address fairness issues. This

framing draws inspiration from the way that checklists are used within the construction industry to prompt and guide critical conversations between stakeholders [10, 13, 28]. Some participants made analogies to accessibility, noting that their teams reviewed the accessibility of their systems at predetermined intervals prior to shipping but did not have analogous reviews for fairness. Conversations about fairness could take place at regularly scheduled product review meetings, with a focus on anticipating fairness issues, or as needed when team members identify fairness issues that need to be addressed. Practitioners speculated that fairness checklists could prompt team members to "*say that we need to have a milestone review*" (P4). Alternatively, as one participant shared, fairness checklists could help data scientists and ML engineers to "*already be thinking about 'Oh, I need to make sure that I'm looking for certain characteristics in the data that may be unfair', flagging it and then sharing that in the scrum meeting with the team when we have a standup*" (P13). This use of checklists is similar to the role that checklists play in the construction industry, where they have enabled "workers to be empowered to 'stop the production line"' [10], and in healthcare, where they have "opened lines of communication and given voice to multiple members of the surgical team" [36], including nurses and surgeons in the operating room [13, 21]. Even in the technology industry, interdisciplinary conversations during development and deployment process are often sites of intervention for value-sensitive design work, as evidenced by Judgment Call [1], Envisioning Cards [11], and more [12, 37, 38]. As one participant shared (see quote in the sidebar), disagreement around ethical concepts may be healthy, but "*these are hard conversations*," and without a framework to enable them, they may not happen on their own.

"To make good AI products, you need data scientists and AI groups that will push things in one direction. And you also need UX people to tell the user story and say 'No' and push back... So these are very healthy checks and balances, but these are hard conversations." (P8)

Fairness and other concepts relating to ethics are situated within particular sociocultural contexts and are therefore often contested. As a result, no checklist can or should play the role of guarding against every possible fairness issue. Indeed, fairness issues are hard to anticipate, and people may disagree about whether particular system behaviors are unfair or not. Many existing checklists intended to help teams anticipate and address ethical issues (e.g., [3, 7, 9, 16, 18, 22, 34]) appear to draw inspiration from checklists in software development, often framing the ethical development and deployment of AI systems as a sequence of discrete technical actions [2, 4, 8, 14, 35]. We argue that a more relevant source of inspiration might be the construction industry, where checklists serve as tools to prompt and guide critical conversations between stakeholders. When framed in this way, fairness checklists could introduce good friction into the development and deployment process, empower practitioners to raise concerns about fairness issues that they might not otherwise feel comfortable raising, and serve as "values levers" in enabling and structuring conversations about fairness that might not otherwise take place.

REFERENCES

[1] Stephanie Ballard, Karen M Chappell, and Kristen Kennedy. 2019. Judgment call the game: Using value sensitive design and design fiction to surface ethical concerns related to technology. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 421–433.

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

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

[4] M Evren Coskun, M Melta Ceylan, Kadir Yigitözu, and Vahid Garousi. 2016. A tool for automated inspection of software design documents and its empirical evaluation in an aviation industry setting. In *2016 IEEE Ninth International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 287–294.

[5] 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.

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

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

[8] 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.

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

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

[11] Batya Friedman and David Hendry. 2012. The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1145–1148.

[12] Batya Friedman, David G Hendry, and Alan Borning. 2017. A survey of value sensitive design methods. *Foundations and Trends in Human-Computer Interaction* 11, 2 (2017), 63–125.

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

[14] 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.

[15] 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, et al. 2009. A surgical safety checklist to reduce morbidity and mortality in a global population. *New England Journal of Medicine* 360, 5 (2009), 491–499.

[16] 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

[17] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. 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. https://doi.org/10.1145/3290605.3300830

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

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

[20] 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.

[21] Lorelei Lingard, Sherry Espin, B Rubin, Sarah Whyte, M Colmenares, GR Baker, Diane Doran, E Grober, B Orser, J Bohnen, et al. 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.

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

[23] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[24] James W Malazita and Korryn Resetar. 2019. Infrastructures of abstraction: how computer science education produces anti-political subjects. *Digital Creativity* 30, 4 (2019), 300–312.

[25] 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.

[26] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 39–48.

[27] Samir Passi and Phoebe Sengers. 2020. Making data science systems work. *Big Data & Society* 7, 2 (2020), 2053951720939605.

[28] D Darshi de Saram and Syed M Ahmed. 2001. Construction coordination activities: What is important and what consumes time. *Journal of Management in Engineering* 17, 4 (2001), 202–213.

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

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

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

[32] Kajsa Simu. 2006. *Risk management in small construction projects*. Ph.D. Dissertation. Luleå tekniska universitet.

[33] Katta Spiel, Emeline Brulé, Christopher Frauenberger, Gilles Bailly, and Geraldine Fitzpatrick. 2018. Micro-ethics for participatory design with marginalised children. In *Proceedings of the 15th Participatory Design Conference: Full Papers-Volume 1*. 1–12.

[34] 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

[35] Michiyo Wakimoto, Shuji Morisaki, and Shuichiro Yamamoto. 2019. A Case Study of Requirements Ambiguities and Goal-oriented Focused Requirements Specification. In *2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 908–913.

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

[37] Richmond Y Wong, Deirdre K Mulligan, Ellen Van Wyk, James Pierce, and John Chuang. 2017. Eliciting values reflections by engaging privacy futures using design workbooks. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–26.

[38] 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. https://doi.org/10.1145/3274463