All the tools, none of the motivation: Organizational culture and barriers to responsible AI work

Amy Heger, Samir Passi, Mihaela Vorvoreanu

As applications of artificial intelligence (AI) have proliferated so too have ethical concerns regarding their potential to cause harm to society. As a result, many organizations that build or use AI systems have developed frameworks or codes of conduct specifying ethical or responsible AI principles they strive to follow (e.g., transparency, fairness, privacy, accountability, safety, etc.; Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020; Greene, Hoffmann, & Stark, 2019; Jobin, Ienca, & Vayena, 2019; Zeng, Lu, & Huangfu, 2019). Although a promising first step, espousing values is not synonymous with acting in accordance with them (Kish-Gephart, Harrison, & Treviño, 2010; McNamara, Smith, & Murphy-Hill, 2018). In fact, many researchers and practitioners have pointed to the inadequacy of such high-level ethical principles without supporting practices to guide on-the-ground implementation (Boddington, 2017; Hagendorff, 2020; Mittelstadt, 2019; Schiff, Rakova, Ayesh, Fanti, & Lennon, 2020). Wide-ranging efforts to bridge this "principles-to-practices gap" have emerged in the form of toolkits, checklists, processes, and evaluative metrics (e.g., Morley Floridi, Kinsey, & Elhalal, 2020). These approaches, however, are overwhelmingly geared towards AI practitioners and place the onus of addressing this gap squarely on individuals (Hagendorff, 2020; Miller & Coldicutt, 2019; Rakova, Yang, Cramer, & Chowdhury, 2021; Stark & Hoffmann, 2019). We believe the current focus on such bottom-up methods, in both industry and academia, is insufficient due to its misalignment with top-down organizational influences.

The culture of an organization in which an AI practitioner is situated, and the priorities of its leadership determine whether the benefits of on-the-ground responsible AI (RAI) tools and best practices can be fully realized (Madaio, Stark, Wortman-Vaughan, & Wallach, 2020; Rakova et al., 2021). In fact, in our own research we found RAI practices were rarely discussed without consideration of the organizational resources and/or incentives necessary for them to be undertaken. We set out to create a RAI maturity model to serve as a map for organizations and AI practitioners to identify where they currently are and where they could go next. To do this, we first conducted 47 semi-structured interviews with RAI consultants and practitioners from different technology companies. We asked about their experiences working with product teams on RAI practices and had them characterize the maturity of these practices.

Despite our objective to learn about less and more mature RAI practices and applications, participants instead wanted to talk about the importance of higher-level organizational factors that first had to be in place before they could even begin to talk about what practices to employ. Several participants indicated that the ability to practice RAI, let alone in a mature way, is contingent on the allocation of resources (e.g., time, budget, headcount, governance, training) and alignment of incentives (e.g., recognition by management, compensation via KPIs and performance evaluations) both of which depended on organizational leadership's prioritization of RAI. For example, they discussed maturity as *"[team leaders] put the resources required behind the team"* (P29) and *"not just saying that responsible AI is important, but actually shifting at a structural level how products are developed, how different kinds of work is incentivized*" (P47). Whereas a lack of maturity was described as "*If the leader doesn't really focus or talk about responsible AI at all … we'll probably see that those teams are not really engaged with any sort of [RAI] activity or tools."* (P34). Overall, qualitative data analysis of this first phase revealed organizational foundations and team approaches as primary categories of maturity in addition to practices. Further, maturity dimensions within these three categories were often interdependent such that progress on one dimension (e.g., practice of mitigating harms) was not feasible until a certain level of maturity was achieved on a different dimension (e.g., leadership and culture prioritizing RAI). The second phase of our research—currently 39 interviews and focus groups with 51 RAI experts to validate

dimensions, fill in gaps, and refine maturity levels—so far has corroborated this multi-faceted and interdependent representation of RAI maturity.

Our RAI maturity model research is not the first to offer robust evidence of organizational culture's crucial role in facilitating the translation of RAI principles into practice. Madaio et al. (2020) found that for effective implementation of their AI fairness checklist, organizational context must be taken into consideration. While AI practitioners saw the checklist as a helpful way to formalize ad hoc processes and empower them to raise concerns, they also stressed the need for such an artifact to be supplemented with organizational processes and infrastructure, aligned with their organization's culture and priorities, and closely tied to performance metrics. Rakova et al. (2021) discovered similar prevalent needs for better organizational processes and structures to assist individuals willing to dedicate their time—often uncompensated and unrecognized—to RAI work practices. Similar to Madaio et al., their participants described a lack of organizational incentives in alignment with RAI practices as a challenge, especially because fast-paced product development and timely shipping tended to be rewarded, despite potential incompatibility with RAI principles. They reported most participants "*see misalignment between individual, team, and organizational level incentives and mission statements within their organization*" (p.16, Rakova et al., 2021). Our research suggests that alignment across these levels is necessary to address the principles-to-practice gap and call for increased research in this area.

Although acknowledged as an impactful factor for the success of responsible AI implementations, we feel the role of organizations—their priorities, leadership, and broader culture—is frequently understated and under-addressed in the RAI community. We are not questioning the valuable contribution of work building out methodologies and tools for practical application of RAI principles. User studies indicate AI practitioners request this guidance, making such efforts critical to continue (e.g., fairness [Holstein, Wortman-Vaughan, Daumé, Dudík, & Wallach, 2020; Madaio et al., 2020], transparency [Heger, Marquis, Vorvoreanu, Wallach, & Wortman-Vaughan, 2022; Mitchell et al., 2018], UX design [Gray & Chivukula, 2019]). We simply posit that for effective widespread adoption of RAI principles the RAI community needs to work on multiple fronts. For example, several toolkits and methods already exist, which can create potential confusion about their utility and place in the AI system development and deployment lifecycle (Morley et al., 2020; Schiff et al., 2020; Vakkuri et al., 2021). One reason we made a RAI maturity model was to provide a framework in which lower-level practices could be organized and understood in the broader context of their organization's ecosystem and team processes. Is trying to examine, design, and improve on-the-ground practices, without giving adequate attention to the environment their application is steeped in, bordering on negligent?

Additionally, work on system-level influences on RAI implementation, such as development of government regulations and the establishment of professional community norms, is invaluable. The fact is, passing regulations is slow and developing norms for a profession that lacks history and standardized mechanisms of accountability is difficult (Mittelstadt, 2019). Due to these and other barriers to system-level approaches, AI practitioners are primarily reliant on their organization to provide the motivation and support for doing RAI work. This is not to say organization-level factors do not face similarly challenging barriers, however, due to a current lack of clearcut external incentives (e.g., industry standards, government regulation, audit requirements), it largely falls on organizations to facilitate RAI.

Difficulty assessing the impact of RAI practices is one reason given for why organizations frequently fail to incentivize this work. Many technology companies base compensation on data-driven performance evaluations and objectives and key results (OKRs) tied to revenue growth. In such

organizations, the question becomes how does one quantify, for example, the benefits of preventing a harmful algorithmic outcome or creating additional model documentation, to ensure proper credit for their work? We question the premise of this approach more broadly: is it appropriate to require that work practices be quantifiable to fit preexisting organizational incentive frameworks? Organizations need to shift to accommodate qualitative assessments and interdisciplinary work in performance evaluations, both of which are core to RAI practice (Rakova et al., 2021). We believe much could be learned from assessment of organizational culture considered in other industries. Nuclear power plant facilities in the United States, for example, assess their organization's safety and security culture every few years. These assessments combine both qualitative and quantitative elements, and like maturity models, help gauge whether an organization's actions align with their espoused principles. Further, entire lines of research in industrial/organizational (I/O) psychology address organizational ethical culture (Treviño, 1990) and ethical climate (Victor & Cullen, 1988), their measurements, and associated outcomes. Analogous to discussions of responsible AI conduct, ethical conduct in organizations is understood to result from a combination of individual characteristics and organization rules and reward systems (Jones, 1991).

We advocate for an increased focus on initiatives and strategies at the organization-level that strive to facilitate prioritization of RAI in organizational culture. The development of our RAI maturity model made abundantly clear that maturity of RAI practices is closely linked to maturity of an organization's resources and incentives. Based on this work and extant research, we believe it is impossible for bottom-up efforts and advocacy, for the most part driven by impassioned individuals, to have nearly the impact as organizational-level initiatives (e.g., culture improvement programs, change management initiatives, senior leadership education). Initiation of such efforts should be undertaken in collaboration with I/O psychologists, business consultants, social psychologists, sociologists, and other experts in relevant fields.

Organizational culture that prioritizes RAI through resources for its implementation in practice (e.g., governance, tools, time, training), and incentives to make use of these resources (recognition, compensation, shared values), is vital if widespread adoption of RAI is to occur. We hope to facilitate discussion at this workshop on why there is such little research on organizational-level efforts to shape culture and senior leadership priorities to align with RAI principles. Is this perhaps due to perceptions of organizational change as being challenging, complex, or simply futile when it comes to the technology industry? Is the RAI community's greater focus on practical guidance, metrics, and tools simply an artifact of their background skewing technical? Is there a vested interest in maintaining the convenient bottom-up framing of RAI as a task that AI practitioners are responsible for doing? We're eager to discuss these topics and many others, in addition to learning about other forms of culture and their influence on AI systems.

REFERENCES

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI* (SSRN Scholarly Paper ID 3518482). Social Science Research Network. https://doi.org/10.2139/ssrn.3518482

Gray, C. M., & Chivukula, S. S. (2019). Ethical Mediation in UX Practice. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3290605.3300408

Greene, D., Hoffmann, A.L., & Stark, L. (2019). Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. *HICSS*.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Heger, A. K., Marquis, L. B., Vorvoreanu, M., Wallach, H., & Wortman Vaughan, J. (2022, November). Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *2022 Conference on Computer Supported Cooperative Work*. https://www.microsoft.com/en-us/research/publication/understanding-machine-learning-practitioners-data-documentation-perceptions-needs-challenges-and-desiderata/

Holstein, K., Vaughan, J. W., Daumé, H., Dudík, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300830

Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat. Mach. Intell. 1*, 389–399.

Jones, T. M. (1991). Ethical Decision Making by Individuals in Organizations: An Issue-Contingent Model. *The Academy of Management Review*, *16*(2), 366–395. https://doi.org/10.2307/258867

Kish-Gephart, J. J., Harrison, D. A. & Treviño, L. K. (2010). Bad apples, bad cases, and bad barrels: Meta-analytic evidence about sources of unethical decisions at work. *J. Appl. Psychol. 95*, 1–31.

Madaio, M. A., Stark, L., Wortman-Vaughan, J., & Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3313831.3376445

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development? *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 729–733. https://doi.org/10.1145/3236024.3264833

Miller, C., & Coldicutt, R. (2019). People, Power and Technology: The Tech Workers' View. *Doteveryone*, 1–24.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2018). Model Cards for Model Reporting. *ArXiv*, *Figure 2*, 220–229.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, *26*(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for shifting Organizational Practices. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–23. https://doi.org/10.1145/3449081

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2020). Principles to Practices for Responsible AI: Closing the Gap. *ArXiv*.

Stark, L., & Hoffmann, A. L. (2019). Data is the new what? Popular metaphors & professional ethics in emerging data culture. *Journal of Cultural Analytics*, 1–22. http://doi.org/10.22148/16.036.

Trevino, L. K. (1990). A Cultural Perspective on Changing and Developing Organizational Ethic. *Research in Organizational Change and Development, 4,* 195–230.

Vakkuri, V., Jantunen, M., Halme, E., Kemell, K.-K., Nguyen-Duc, A., Mikkonen, T., & Abrahamsson, P. (2021). Time for AI (Ethics) Maturity Model Is Now. *ArXiv:2101.12701 [Cs]*. http://arxiv.org/abs/2101.12701

Victor, B., & Cullen, J. B. (1988). The Organizational Bases of Ethical Work Climates. *Administrative Science Quarterly*, *33*(1), 101. https://doi.org/10.2307/2392857

Zeng, Y., Lu, E., & Huangfu, C. (2018). *Linking Artificial Intelligence Principles*. 5. http://arxiv.org/abs/1812.04814.