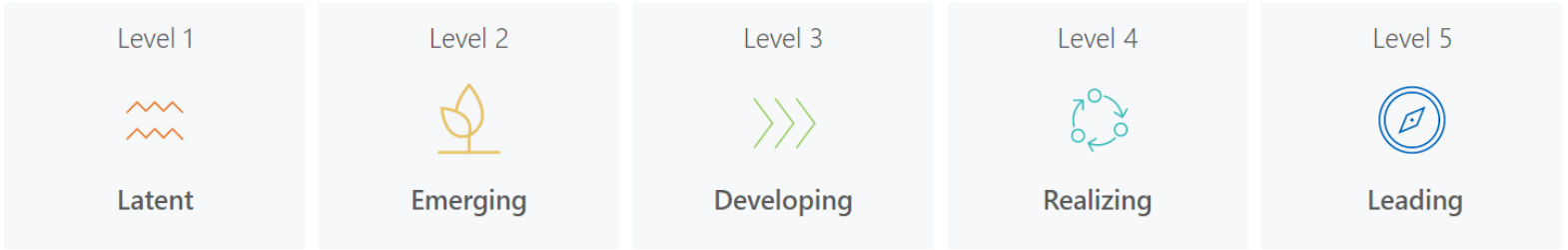
RESPONSIBLE AI MATURITY MODEL

MAPPING YOUR ORGANIZATION'S GOALS ON THE PATH TO RESPONSIBLE AI



MIHAELA VORVOREANU ▪ AMY HEGER ▪ SAMIR PASSI ▪ SHIPI DHANORKAR ▪ ZOE KAHN ▪ RUOTONG WANG

AETHER CENTRAL UX RESEARCH & EDUCATION ▪ MICROSOFT

V1 ▪ MAY 17, 2023

Executive summary

The Responsible AI Maturity Model (RAI MM) is a framework to help organizations identify their current and desired levels of RAI maturity.

The RAI MM contains 24 empirically derived dimensions that are key to an organization's RAI maturity. The dimensions and their levels are based on interviews and focus groups with over 90 RAI specialists (e.g., RAI champs, MSR experts) and AI practitioners (e.g., user experience (UX) researchers, UX designers, data scientists). Each dimension has five levels, going from low (Level 1: Latent) to high (Level 5: Leading) maturity. The dimensions are organized into three main categories:

- Organizational Foundations
- Team Approach
- RAI Practice

The Responsible AI Maturity Model is a map to the complex territory of RAI.

We recommend thinking of the RAI MM as a high-level map of the complex and evolving territory of RAI. Use it to help you navigate what it means to be a mature RAI organization. Looking ahead at higher maturity levels to see what is possible and desirable is more important than assessing your organization's or team's current level. As a field, RAI is not yet mature, so expect your organization or team's maturity level to reflect that.

You're using the RAI MM level "right" if you allow it to catalyze learning about RAI and discussions on organizational change.

Acknowledgments

We thank the following individuals who went above and beyond to support this work: Steph Ballard, Michael Benton, Nokta Berberoglu, Dean Carignan, Neil Coles, Miro Dudik, Ruth Kikin-Gil, Daniel Kluttz, Bernhard Kohlmeier, Kim Laine, Kristen Laird, Michael Madaio, David Marcos, Hugh North, Besmira Nushi, Jen Sillik, Sridhar Sriram, Forough Poursabzi, Victor Poznanski, Sudipto Rakshit, Kelley Rand, Mehrnoosh Sameki, Siddhartha Sen, Robert Sim, Maya Smooth, Hiwot Tesfaye, Anja Thieme, Kathleen Walker, Hanna Wallach, Jenn Wortman-Vaughan, Ben Zorn.

Table of contents

Introduction to the Responsible AI Maturity Model

Why do we need the RAI Maturity Model?

We are responsible for the AI systems we develop. Government regulation of AI systems is forthcoming. But, until policies are in place, the onus for developing and deploying AI systems responsibly falls entirely on the organizations that build and use the technologies. As a result, adoption of RAI principles by companies has surged as a form of self-regulation. Yet research shows effectively translating these principles into practice is challenging (Mittelstadt, 2019; Sanderson et al., 2022; Schiff, Rakova, Ayesh, Fanti, & Lennon, 2020). To overcome this challenge, wide-ranging efforts have emerged in the form of toolkits, checklists, practical guidance, and metrics. These approaches, however, are often geared towards individual AI practitioners and not organizations. Such efforts are often piecemeal, lacking clarity on how they fit into an organization's larger RAI strategy. Without consensus on many best practices or what it means to be mature yet, we recognize a need to assess and map this uncharted new territory. The RAI MM is an important next step and does just that—it identifies the core components of RAI maturity for an organization and how they fit together.

How was the RAI Maturity Model developed?

The RAI MM identifies the core components of RAI maturity and how they fit together.

We looked at the research literature and identified the need for an RAI maturity model. We found no maturity models specific to RAI. The maturity models we found on AI/ML insufficiently addressed RAI (e.g., Alsheibani et al., 2019, IBM, 2021, Keystone.ai, 2021, Oracle, 2020, Ovum, 2018), demonstrating a clear need for an RAI-specific maturity model in alignment with recent calls for such an artifact (e.g., Shneiderman, 2020, Vakkuri, et al., 2021). We scoped the RAI MM to only include RAI topics due to preexisting maturity models that cover general AI topics. [*See the appendix for references to related maturity models.*]

We created the RAI MM using a rigorous two-step research process informed by best practices for building maturity models (Becker et al., 2009; De Bruin et al., 2005; Mettler, 2011). The RAI MM is the

result of more than 80 hours of interviews and focus groups with a total of 90 participants, and hundreds of hours of analysis and synthesis.

We conducted 47 interviews with Microsoft internal and external RAI specialists who work across product teams and AI practitioners who work within product teams to better understand what factors contribute to RAI maturity. We asked participants to describe variation in RAI practices they had seen or performed, specifying those that they perceived as more or less mature. We then coded and analyzed the interview data to create an initial draft of RAI maturity dimensions.

We engaged 56 internal experts in an iterative creation process to further build out the maturity model. Across 23 focus groups, 17 interviews, and additional asynchronous feedback, these experts helped to validate the RAI MM dimensions, identify gaps, and develop discrete maturity levels for each dimension.

The next step is to pilot the RAI MM with teams inside Microsoft. Contact the authors if you'd like to participate or have any feedback.

What is important to know about the RAI maturity model?

Collaboration is central to RAI maturity.

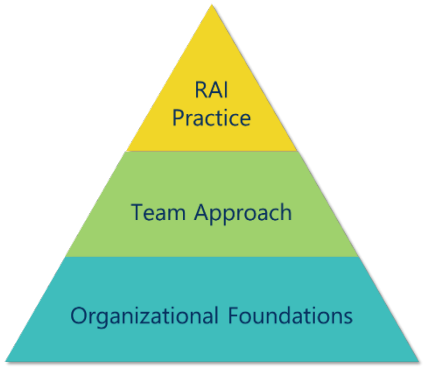
Collaboration surfaced as the core driver of RAI maturity across all 90 people we consulted. Collaboration among disciplines, roles, product teams, and RAI specialists is at the center of what it means to do RAI in a mature way. RAI does not have simple solutions. It requires hard conversations and consideration of trade-offs. For a team to even be aware of trade-offs, they need multiple perspectives in the room.

As a practice that is not technical or easily quantified, collaboration work often remains invisible; it's overlooked and undervalued despite its fundamental role in organizational maturity. Collaborative engagements—when and if they happen, who's involved in them, if they critically anticipate or address RAI risks—are such an essential aspect of RAI work that no organization should expect to reach high levels of maturity without first working to build out maturity of cross-discipline collaborations.

RAI maturity dimensions are interdependent.

Our research shows that dimensions in the maturity model are highly interdependent: progress on one dimension often depends on the levels of maturity of other dimensions. We have created discrete dimensions to facilitate ease of understanding and assessment. However, such clean separation of concepts is a simplification of the complexity of the RAI landscape. Dimensions in the three main categories depend on each other, as follows:

- ***Organizational Foundations*** are the dimensions of RAI maturity that pertain to the organization as a whole and are directly impacted by decisions of the senior leadership team. As the foundation of the pyramid, these dimensions lay the necessary groundwork for mature growth in the other two categories: *Team Approach* and *RAI Practice*.
- ***Team Approach*** are the dimensions of RAI maturity that pertain to the way teams approach RAI work (how, why, with whom). As the center of the pyramid, these dimensions address how people work as they engage in *RAI Practice* and depend on *Organizational Foundations*.
- ***RAI Practice*** are the dimensions of RAI maturity that pertain to how teams perform specific RAI work such as identifying, measuring, and mitigating RAI risks. These dimensions depend on maturity in the other two categories: *Organizational Foundations* and *Team Approach*.



Crucial considerations of the RAI Maturity Model

The RAI MM will continue to evolve.

- RAI is a new and constantly shifting field with a lot of unknowns. We are just beginning to understand RAI best practices. As a result, this maturity model is a living artifact and will continue to evolve.

The RAI MM is forward-looking.

- The RAI MM captures the ideal goal-state of RAI, rather than its current state. Because RAI is newly developing as a practice in most organizations, we think an aspirational model provides the most informative maturity guidance. This also means most organizations currently are at lower levels of maturity and level 5 is highly aspirational.

RAI maturity may vary within your organization.

- Differences in RAI maturity will likely exist across your organization. For example, some teams may operate at higher levels of maturity than others particularly on the *Team Approach* or *RAI Practice* dimensions. Although levels of *Team Approach* and *RAI Practice* dimensions can be averaged across teams for an understanding of the organization's overall maturity, consideration of team differences will help focus improvement efforts, and allow those higher in maturity to share successful practices and lessons learned.

Progress across RAI maturity levels varies in difficulty (i.e., progression is characterized by an exponential, not linear curve).

- Although the maturity levels are labeled with values 1 (latent) through 5 (leading), they are not equivalently spaced incremental steps. The effort necessary to move from one maturity level to the next varies, with more significant work often required to move between lower levels. Advancement out of levels 1 (latent) and 2 (emerging) can require creating new processes or practices, which is resource-intensive, whereas progress out of levels 3 (developing) and 4 (realizing) might just be formalizing or integrating existing processes or practices, a somewhat easier lift. Also, keep in mind that because dimensions are connected and contingent, progress

across levels might be challenging or restricted due to a lack of maturity of a different dimension.

RAI work is context-specific.

- RAI work is context dependent and is determined, in large part, by the type and domain of the AI system. RAI practices thus can vary considerably. This creates difficulties when trying to develop concrete yet generalizable criteria for how teams should *do* RAI practices on-the-ground. It is possible to define characteristics of how to *approach* RAI practices in a mature way—what discussions must happen, what sociotechnical factors must be considered, what approaches must be taken, and what decisions must be made.

How to use the RAI Maturity Model

Use the RAI MM as a map for guidance, not as a measurement tool for punitive purposes.

- The maturity model is a framework that organizes the complex territory of RAI. It is a roadmap of maturity progression so organizations and teams can identify where they are and where they could go next. Identify what level you want to aim for and discuss what it would take to get there. Keep in mind that level 5 might not be applicable or desirable for every organization.

Use the RAI MM as a map for guidance, not as a measurement tool for punitive purposes.

Keep in mind the context of each dimension, and not average scores across dimensions.

- Calculating an overall maturity score or comparing scores across dimensions is not recommended. This is an inappropriate usage of the RAI MM because some maturity dimensions are more impactful than others or are reliant on others, and therefore progress on them can only be made after another dimension reaches higher maturity. For example, a level 5 in *tooling* does not have the same impact as level 5 in *culture and leadership*. Therefore, a particular high level is not meaningful when abstracted away from the context of its dimension and interdependency with other dimensions.

Use the RAI MM to supplement, not replace, other maturity models (AI, security, privacy).

- The RAI MM is not a comprehensive maturity model that covers all of AI, data or technology within an organization. Use it in concert with other relevant maturity models.

Have both senior leaders and team members reflect on the organization's RAI maturity. Note any discrepant perceptions of maturity.

- Individuals in particular positions may be better suited to accurately reflect and assess on particular categories of dimensions—teams about their approach and practice and leadership and RAI specialists about organizational priorities and capacity. Even so, all users of the maturity model should assess all dimensions, as leaders and teams may vary in their perceptions. Use divergence of perceptions as an important signal that there is room for higher maturity on a given dimension.

Ensure you satisfy all criteria of one maturity level before moving on to the next level.

- To progress up a level in maturity, you must first achieve any desirable goals or positive facets of the previous level. For example, it is assumed that if level 4 says "teams are able to identify RAI risks effectively" that those at level 5 are also able to do so as well because they first had to satisfy the criteria for level 4 before reaching level 5.

Dimensions of the RAI Maturity Model

Organizational foundations

What: *Organizational Foundations* are the dimensions of RAI maturity that pertain to the organization, or company, as a whole. These dimensions are directly impacted by decisions of the senior leadership team, who often are the only ones in position to catalyze progress on them. These dimensions are foundational to an organization's ability to achieve RAI maturity and often must be in place for progress to occur on *Team Approach* and *RAI Practice* dimensions.

Organizational Foundations dimensions are further divided into:

- *Leadership and Culture*
- *Organizational Capacity*, which includes:
    - *Governance*
    - *Knowledge Resources*
    - *Tooling*.

Who: *Organizational Foundations* dimensions should be assessed by both the organization's senior leadership team and teams to understand how leadership decisions are experienced on the ground. Potential disparities are great signals for growth opportunities.

Leadership and Culture

Leadership and Culture reflects how much an organization's leadership and culture prioritize RAI—to what extent RAI values are translated into resources and incentives for RAI work. Organizational leaders' actions highlight their priorities, signaling to employees what is valued and rewarded by the organization. Leadership's prioritization of RAI in their decision-making empowers teams to approach and practice RAI maturely.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
The organization might have some general messaging about RAI, but… • C-suite leaders do not consider RAI in decision-making or allocate resources to it. For example, budget is not set aside for tools, consulting with experts, education, collecting representative data, and headcount. • The organization does not incentivize RAI work. • Management does not buy-in to the importance of RAI, as demonstrated by a lack of change in decisions and behavior to align with RAI values.	The organization states it values RAI, (e.g., creates messaging, establishes RAI principles), but… • C-suite leaders do not consider RAI in decision-making or allocate resources (budget, headcount) to it. • The organization does not incentivize RAI work. • Management does not allocate time for RAI practices during the product development and deployment lifecycle. • RAI work, if done, is driven by a few passionate individuals.	The organization states it values RAI, (e.g., creates messaging, establishes RAI principles), and… • C-suite leaders prioritize some RAI aspects (e.g., fairness, transparency) but not others (e.g., reliability and safety, accountability). • Some resources are allocated to jumpstart RAI efforts. • The organization encourages RAI work but does not incentivize it. • Management points to competing values and shipping pressures as a justification for not prioritizing RAI. • RAI work is driven by passionate teams in a few pockets of the organization.	RAI is valued and prioritized by the organization. C-suite leaders prioritize RAI in their decision-making and resource allocation (e.g., investments in training, governance, infrastructure, RAI experts). Organization plans long-term investment in RAI - headcount & budget for RAI practice, research & education. Some teams are incentivized to prioritize RAI, possibly with explicit RAI commitments, performance indicators, and management recognition.	RAI is fully integrated in the organization; it is part of business as usual and a required aspect of AI systems. AI products are not released if at least major known RAI risks (e.g., fairness-related harms) are not addressed. C-suite leaders prioritize RAI in their decision-making and resource allocation. They set clear RAI commitments and performance indicators. The organization incentivizes all AI teams to prioritize RAI.

Organizational capacity

Organizational capacity includes foundational dimensions that afford teams the ability to do RAI work: *governance*, *knowledge resources*, and *tooling*. These dimensions are often interdependent with one fostering progress in another—*RAI policy* can facilitate creation of *knowledge resources*, *knowledge resources* can aid *tooling*, *tooling* can enable RAI compliance processes and infrastructure.

Governance

Governance refers to the existence of formal organizational policies (*RAI Policy*, e.g. Microsoft's RAI Standard), and the infrastructure and practices needed to facilitate compliance with those policies across the organization (*RAI Compliance processes and infrastructure*). It provides teams the guidance, processes, and systems needed to implement *RAI Practice* dimensions, and enables maturity on the *Accountability* dimension.

RAI policy

RAI policy refers to compliance requirements and guidelines for an organization's AI products. A well-defined RAI Policy enables accountability within an organization by formalizing requirements, delivering a reference point to identify when teams and the organization deviate from desired actions. Furthermore, *RAI Policy* provides a good starting point for organizations who have few RAI practices in place.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
The organization does not have any internal RAI policies.	The organization is in the process of creating a unified internal RAI policy. Some RAI policies specific to some RAI aspects (e.g., fairness, privacy) might exist, but they are fractured.	The organization has an internal RAI policy. The policy is supported by some additional documentation explaining the policy and how to implement it. However, the documentation is high-level, and does not cover all necessary aspects.	The organization's internal RAI policy is rigorous (i.e., based on empirical evidence and research), comprehensive, and auditable. RAI policy is continuously updated & improved based on lessons learned from its application. The policy is supported by comprehensive documentation, but it might still be missing some in-depth aspects needed for implementing it (e.g., detailed practices on how to measure some types of RAI harms).	The organization's internal RAI policy is so forward-thinking that it informs external policies (e.g., standards, laws, regulations). The organization's RAI policy is reconciled against the entirety of a company's other internal policies an governance processes (e.g., RAI, pri or security policies and processes ar integrated, and redundancies are removed). The policy is supported by documentation that is integrated into processes and tools. Processes exist for keeping this documentation up to date and incorporating new knowledge over time.

RAI compliance processes and infrastructure

RAI compliance processes and infrastructure refers to the practices, procedures, and internal structures set up to facilitate compliance with *RAI policy* across the organization. Examples of *RAI compliance processes and infrastructure* include guidance on doing risk assessments (e.g., impact assessments), frameworks/templates for filling out required documentation (e.g., transparency notes) and processes for reporting and reviewing high risk uses of AI.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
The organization does not have a RAI-specific compliance process or infrastructure in place. For instance, there is no structured process in place for employees to report RAI harms.	RAI compliance processes are in planning. Any RAI compliance work (e.g., RAI reviews) is done manually, in an ad-hoc manner, without infrastructure. For instance, the organization has an informal process for reporting RAI harms.	RAI compliance processes vary a lot across organizational units. Infrastructure for RAI compliance is fragmented. For instance, the organization has a structured process for reporting RAI harms (e.g., company-wide process for high-risk uses).	RAI compliance processes are streamlined across the organization. Infrastructure for RAI compliance exists but is disconnected from product teams' AI development and deployment lifecycle. For instance, the organization has a structured process for reporting RAI harms and for disseminating information about the harms to relevant personnel. There are processes in place to address RAI harms on a case-by-case basis.	The organization's RAI compliance processes are scalable, extensible, and continuously improved. Infrastructure for RAI compliance is deeply integrated into product teams' AI development and deployment lifecycle. Automation is used where appropriate. The organization has a structured process for reporting RAI harms, which is integrated into teams' AI development and deployment lifecycle. The organization has a fully functioning RAI response process to systematically address different types of identified RAI harms.

Knowledge resources

Knowledge resources refers to the availability of RAI experts and RAI-specific knowledge, training, and education resources within an organization. Knowledge resources are fundamental for gaining maturity in almost every other dimension and are the primary means of educating practitioners and teams about RAI principles and practices.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
Training resources specific to the organization's needs are non-existent or limited. A lot of RAI knowledge is undocumented. Access to RAI experts is limited.	Training resources specific to the organization's needs are non-existent or limited. A lot of RAI knowledge is undocumented. Teams have access to some, but not sufficient, RAI experts. Pockets of RAI expertise begin to emerge bottom-up within the organization.	Some of the organization's RAI knowledge is codified in training resources, but these are incomplete and/or not actionable (e.g., too generic, or hard to understand/apply). Teams have access to in-house RAI experts.	The organization's RAI knowledge is codified in best practices, guidance, and effective training (e.g., case studies, cross-disciplinary workshops). The organization maintains an up-to-date library of actionable guidance for addressing RAI issues. RAI experts are part of a recognized external community that generates and shares knowledge.	The organization abstracts and generalizes from its own learnings into knowledge that can be used by others. The organization's RAI experts engage publicly in knowledge creation and sharing (e.g., by participating in professional and academic conferences).

Tooling

Tooling refers to an organization's access to and development of tools that facilitate RAI work. Tools are invaluable for scaling up RAI work in an organization because they support implementation of RAI practices on-the-ground (e.g., identifying and measuring RAI risks) in service of high-level RAI principles (e.g., reliability and safety, fairness). Note that the mere existence of tools is not enough for RAI maturity without the necessary incentives and compliance requirements for practitioners to use those tools. Factors that influence maturity of tooling include the extent of tooling assistance throughout the AI lifecycle, extent of integration in workflows, and diversity of capabilities and customizability.

[*For links to Microsoft-specific RAI tools, please see the Appendix or https://aka.ms/rai, HAX Toolkit, and RAI Toolbox.*]

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
The organization does not provide AI teams with access to RAI-specific tools.	The organization provides AI teams with access to some RAI-specific tools, but they are: • Fragmented, not inter-operable. • Only capable of addressing limited RAI aspects (e.g., fairness) and stages in the AI development and deployment lifecycle (e.g., model testing).	The organization provides AI teams with access to some RAI-specific tools, but they: • Are not integrated into teams' workflows and infrastructure. • Can't handle the data types or scale of some AI models.	The organization provides AI teams with access to some RAI-specific tools, and they: • Are integrated into teams' processes, customizable, and provide end-to-end support. • Have centralized data types and location, common infrastructure, and automation capabilities. • Handle many scenarios.	The RAI tooling ecosystem is extensible and customizable. AI teams can contribute new tools. Tools support end-to-end documentation and auditing.

Team Approach

What: *Team Approach* are dimensions of RAI maturity that pertain to the way teams approach RAI work (e.g., how, why, with whom). These dimensions are fundamental to an organization's ability to do RAI work and often must be in place for meaningful progress to be made in the *RAI Practices* dimensions. *Team Approach* is also impacted by high-level organizational factors such as *Organizational Capacity* and *Leadership & Culture: Prioritizing RAI,* and thus the dimensions below must be considered in the context of these broader organizational aspects. A large portion of *Team Approach* dimensions fall under cross-discipline collaboration. This focus is further divided into specific elements of collaboration (e.g., motivation, timing) and specific collaborative relationships (e.g., within teams and RAI specialists working with product teams).

Who: *Team Approach* dimensions should be assessed by an organization's teams, as the maturity of such dimensions mostly reflects teams' actions and decisions. Combining team assessments across a department can give a snapshot of the department's average maturity for each team approach dimension (do not average across dimensions). Note that teams within a department can vary drastically in their maturity, thus the focus on efforts to increase maturity in Team approach dimensions should be specifically tailored to teams.

For a faster, but less robust, assessment of *Team Approach* dimensions, the head of a department may assess what level they believe teams in their department fall. These dimensions, however, are less suitable for a leader to assess, because they may not witness some of these dimensions firsthand—e.g., *within team collaboration* and *UX practitioners' AI readiness*.

Teams valuing RAI

Teams valuing RAI refers to the extent to which teams see the necessity and importance of RAI work. Without meaningful buy-in from a team—belief that RAI is critical for developing high quality AI products—there is little likelihood of reaching RAI maturity. RAI policy and compliance processes often require teams to do RAI work. However, motivation is needed to go beyond doing these requirements at the bare minimum level. An organization can have all the Organizational Capacity needed in place for RAI maturity, but if its teams aren't willing to use these resources, little progress will be made.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
Teams believe RAI is an impediment that lowers the quality of AI products. Or, they are simply unaware of/ surprised to learn about RAI work.	Teams are aware of the existence of RAI but see RAI work as a tax that adds little value to AI products.	Teams begin to see some value in RAI work and how it benefits the quality of AI products.	Teams believe in the value of RAI and actively learn and refine their RAI knowledge.	Teams have a deep appreciation of RAI, constantly strive to implement it across all stages of the AI lifecycle, and regularly advocate for it.

Timing of RAI within the AI development and deployment lifecycle

Timing of RAI refers to how early or late during the AI product development and deployment lifecycle teams do RAI work. If RAI concerns are ignored until the end of product development, it's too late to make substantial changes to appropriately mitigate RAI risks. RAI concerns thus must be considered from the onset of the AI lifecycle to inform the design and development of the AI system from the very start. This helps to, among other things, minimize duplication of work and reduce time waste.

Level 1	Level 2	Level 3	Level 4	Level 5
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
Teams do not dedicate time to RAI during the AI development and deployment lifecycle. Teams face RAI concerns or requirements very late, when they thought the work was done.	Teams do not consider RAI early in the AI development and deployment lifecycle. Teams dedicate time to RAI sometime during the AI development and deployment lifecycle, but not at the beginning. This can result in time wasted and a lot of work needing to be redone.	Teams consider RAI early in the AI development and deployment lifecycle, but they do not bring in UX disciplines early.	Teams consider RAI at the start of the AI development and deployment lifecycle. They start UX work early, prior to AI model development.	Teams consider RAI at early ideation stages before they have a product idea and before the product development lifecycle kicks off. Prior UX work informs product idea. This saves time and the need to redo work.

Motivation for AI products

Motivation for AI products refers to the inspiration and rationale behind their development. Many AI products are often built with a desire to implement shiny new AI technologies (technology-centered development) instead of being motivated by user needs (people-centered development). Mature RAI product development is centered around people and how best to address their needs, including paying attention to how the product will impact society and environment. At times, the most mature action from an RAI perspective is to not build the system.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
Technology-centered: Product development prioritizes or is motivated by the existence of a new technology (e.g., the release of a new large model) or pressure to integrate AI into product, not by a demonstrated need people have.	Product development is primarily motivated by technology but considers people at a late stage or in a superficial manner that makes it difficult to identify and address RAI issues. If RAI issues are identified, addressing them is delayed indefinitely or until a future release.	Product development is primarily motivated by technology but considers people at a late stage or in a superficial manner that makes it difficult to identify and address RAI issues. If RAI issues are identified, shipping is delayed to address at least major ones before releasing the product.	**People-centered:** Product development is primarily motivated by people's needs and reflects careful trade-offs among business, experience, technology, and RAI priorities. The outcome of careful deliberations may be a decision not to build if harms outweigh benefits.	Product development is grounded in inclusive, equitable and potentially innovative approaches to understanding people's needs and considering a technology's impact on people, society, the environment, etc. (e.g., approaches such as inclusive design, value-centered design, participatory design, design justice).

Cross-discipline collaboration

Cross-discipline collaboration refers to the people in different roles or with different expertise working together to address RAI problems. Cross-discipline collaboration is indispensable for doing RAI work, because RAI is not a technical problem with solely technical solutions. RAI is about impacts on people and society and therefore requires involvement with experts who understand people, society, and the AI system's application context.

Sociotechnical approach

Sociotechnical approach refers to the extent to which teams approach RAI from a sociotechnical—as opposed to a solely technical—perspective. This impacts what forms of RAI work are considered necessary, how they get done, by whom, and when.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
RAI is viewed as a technical problem, with technical solutions. Technical experts such as data science, engineering, and project management, work alone. They do not consider social aspects or bring in sociotechnical experts such as UX researchers or designers.	RAI is viewed as a technical problem, with technical solutions. Technical experts such as data science, engineering, project management, work alone. They do not consider social aspects. They involve sociotechnical experts, such as UX practitioners, very late in the process and only for superficial work that is not necessarily sociotechnical in nature (e.g., aesthetics of the user interface).	RAI is viewed largely as a technical problem but with some social implications. Technical experts do perfunctory sociotechnical work themselves, without involving sociotechnical experts (e.g., data science, project management do sociotechnical work without engaging UX disciplines).	RAI issues are viewed as sociotechnical problems. Disciplines with sociotechnical expertise such as UX, anthropology, sociology, linguistics, etc. are engaged to address RAI issues.	RAI is viewed as a sociotechnical practice. Disciplines with sociotechnical expertise such as UX, anthropology, sociology, linguistics, etc. are engaged from the very beginning and shape product strategy (e.g., whether and what to build).

Common language

Common language refers to different practitioners, experts, and stakeholders involved in AI product development being aware that they might not have a shared vocabulary and being purposeful about building one. Working across disciplines is difficult because each discipline has its own technical jargon. For example, a *feature* means something different to a data scientist (input variable in a model) than it does to a UX designer or PM (a function or capability of a product). Learning about disciplines other than one's own and developing a common language is necessary for collaboration.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
AI practitioners in different disciplines (e.g., data science, project management, UX) don't have a common language. They lack the ability to understand the vocabulary of other disciplines, stakeholders, customers, RAI consultants, and experts and are not aware they are misunderstanding each other.	AI practitioners in different disciplines cannot communicate effectively but are able to recognize they lack a common language and that they're not on the same page.	AI practitioners in different disciplines have enough common language to understand each other a little bit and to recognize what they don't know. A few disciplines might be able to communicate well (e.g., UX and front-end software development) yet there is little understanding across all disciplines.	AI practitioners in different disciplines have enough of a common language to have useful conversations. They actively seek clarification and better understanding. They invest patience and curiosity in getting on the same page.	AI practitioners in different disciplines have the ability to engage in cross-disciplinary dialogue – to translate from one discipline, stakeholder group, or customer to another and engage in deep meaningful RAI work. The cross-disciplinary dialogue respects each discipline's expertise and does not imply that everyone needs to acquire deep expertise in other areas.

Collaboration within teams

Collaboration within teams refers to the interactions between members of a product or service team when they do RAI work. Collaboration is often taken for granted, but teams need to reflect on how they work and be intentional about creating a team culture that enables them to work together well.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
Teams lack a diversity of perspectives (backgrounds, experiences). Disciplines are in separate siloes and they do not talk to each other. Disciplines are purposefully kept separate for (invalid) reasons such as "saving them time." Team members do not have a holistic view of the product, its rationale, and the work of disciplines other than their own.	Disciplines are in separate siloes but they talk to each other some – for example, only at milestones, integration points (pitching things over the wall), or when a problem occurs. Team members do not have a holistic view of the product, its rationale, and the work of disciplines other than their own.	Different disciplines and diverse perspectives are together in the room, but some disciplines are not heard. Team members have a partial view of the product and its rationale and some understanding of other disciplines' work.	Different disciplines and diverse perspectives discuss and make decisions together. Each team member has a holistic view of the product, its rationale, and the work of disciplines other than their own.	Different disciplines & diverse perspectives work together from the very beginning and on the most important decisions (e.g., what to build, roadmap, resource allocation). Each team member has a holistic view of the product, its rationale, and the work of disciplines other than their own, **and** influence other disciplines. Team members understand how work in their discipline impacts work in other disciplines and know when to seek out other disciplines' expertise.

Non-UX disciplines' perceptions of UX

Non-UX disciplines' perceptions of UX refers to how practitioners of non-UX disciplines (e.g., software engineers, project managers, data scientists) perceive the role of UX disciplines (e.g., research, design, content strategy, copywriting). Because RAI is a sociotechnical problem requiring a sociotechnical approach, the involvement of UX disciplines, who specialize in understanding and designing for people and society, is crucial. Engagement with UX disciplines can be immature if other disciplines do not understand or value them. This dimension influences several other dimensions' maturity: *Timing of RAI, Sociotechnical Approach, and Motivation for AI products*. In turn, maturity on this dimension is influenced by UX practitioners' AI readiness. When considering this dimension, pay attention to discrepancies between UX and non-UX disciplines' perceptions of maturity. Such discrepancies signal opportunities for growth.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
Non-UX disciplines such as data science, engineering, project management, etc. do not understand the value of UX and do not engage with the many UX disciplines (e.g., research, design, writing, content strategy, information architecture) in AI projects.	Non-UX disciplines have a narrow understanding of what UX disciplines do. They might see the role of UX as delivering reductive, simple deliverables. For example, non-UX disciplines may erroneously assume that the role of UX design is merely to deliver wireframes and mock-ups, and the role of UX research is just to conduct usability testing or validate existing assumptions. UX disciplines are not engaged meaningfully – e.g., in research, product strategy, information architecture—to	Non-UX disciplines perceive UX disciplines as useful and see a role for UX, but they perceive work done by UX practitioners as time-consuming, so they do UX work themselves, believing they can do it faster.	Non-UX disciplines perceive UX disciplines as valuable and see a broader role for UX, including bringing new qualitative insights, raising new questions, and identifying risks. These and the resulting changes, even substantial ones, are embraced by non-UX practitioners.	Non-UX disciplines perceive UX disciplines as equal and indispensable. They understand that the role of UX includes shaping strategy, realizing moral and social values beyond the needs of immediate users and stakeholders. Non-UX practitioners understand and respect UX work and how to best collaborate.

	bring new insights or raise new questions. UX practitioners have to advocate and educate others about UX disciplines because, for instance, non-UX disciplines may mistakenly perceive qualitative research as anecdotal, not quantifiable, and therefore lacking value.			

UX practitioners' AI readiness

UX practitioners' AI readiness refers to how knowledgeable and prepared practitioners in UX disciplines (e.g., research, design, and content strategy) are to work with AI. UX practitioners' AI readiness depends on maturity in dimensions such as *Knowledge resources, Collaboration within teams* and *Teams working with RAI specialists*. Maturity on this dimension influences maturity on other dimensions, such as *Sociotechnical approach* and *Non-UX disciplines' perceptions of UX*.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
UX practitioners approach AI as if it were a traditional, deterministic system. UX practitioners leverage traditional UX methods and tools without knowing whether the uniqueness of AI might make them inaccurate or where they might fall short.	UX practitioners appreciate the uniqueness of AI systems but might not know exactly how to approach it. UX practitioners do not assume that traditional methods and tools are appropriate for UX of AI and recognize the need to learn about AI and UX of AI.	UX practitioners approach AI with some understanding of its probabilistic nature. UX practitioners use methods and tools specific to AI, but in an ad-hoc, non-systematic way. UX practitioners might overlook some aspects of probabilistic systems such as failures and, for example, only design for the golden/hero path.	UX practitioners have a system-level understanding of AI, including the UX implications of data science work such as training and testing datasets, data collection methods, the nature and choices related to the ML model, and its explainability. UX practitioners' use of methods and tools specific to AI is fully integrated in their workflows.	UX practitioners approach UX of AI in a systematic, consistent way – e.g., they use AI design frameworks such as the HAX Toolkit. UX practitioners innovate upon UX of AI practices, guidance, methods, and tools.

RAI specialists working with product teams

RAI specialists working with product teams refers to how RAI specialists approach working with product or service teams when advising them. RAI specialists (experts, consultants) need to set the right expectations concerning roles and responsibilities at the start of a collaboration because their initial interactions with teams drastically impact a team's receptiveness to RAI work.

The *RAI specialists working with product teams* dimension should be independently assessed by both parties and discussed together afterwards. If disagreements arise, this is an opportunity for the team and RAI specialists to share what's working and what's not, and together develop an improvement plan.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
RAI specialists set wrong or unrealistic expectations of the engagement – e.g., we'll make your system fair in one week. RAI specialists are alarmist about a product's RAI risks. They react to a team's AI system (design or usage) with many and/or dire concerns. These concerns might lead teams to panic or fear that they need to mitigate all potential RAI risks before they can ship.	RAI specialists set unclear expectations for who is responsible for what and outcomes of the engagement. As a result, teams might either expect RAI specialists to do RAI work for them or assume that all the responsibility falls solely on the team. RAI specialists overwhelm teams with concerns. They make too many recommendations, and share information without distilling it down so teams can reasonably understand and digest it. As a result, teams might give up, feeling like too much needs to be done to address RAI concerns.	RAI specialists set clear expectations about their expertise, intention, nature of the collaboration, outcomes of the collaboration, and responsibilities. Despite this, they only consider teams' RAI work, failing to account for other external pressures that might impact the engagement. RAI specialists show empathy for teams and voice recognition that RAI is new and difficult for all.	RAI specialists set clear expectations about their expertise, intention, nature of the collaboration, outcomes of the collaboration, and responsibilities. They also consider teams' other external pressures that might impact the engagement – e.g., they coordinate with other applicable compliance areas and take into consideration the teams' other priorities & deadlines. RAI specialists are mindful of a team's RAI knowledge and experience and take the appropriate steps to make things manageable – e.g., they are concise, designate one point of contact, and clarify that the team doesn't have to address every single RAI risk.	RAI specialists document expectations and work agreements, and revisit and revise them over time as needed. RAI specialists interact with teams like they are in a partnership; we're all on the same side and working together towards the common goal of making a better product. They frame RAI engagements as a journey throughout the product's lifecycle, not as a one-time effort.

Teams working with RAI specialists

Teams working with RAI specialists refers to how product teams interact with RAI specialists who are advising them.

The *Teams working with RAI specialists* dimension should be independently assessed by both parties and discussed together afterwards. If disagreements arise, this is an opportunity for the team and RAI specialists to share what's working and what's not, and together develop an improvement plan.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
Teams are adversarial in their engagements with RAI consultants/experts or do not engage at all. For example, when teams receive RAI Specialists' input, they might not listen to it, and when teams provide RAI Specialists information, they might share little and not give access to the people on the team with the knowledge the RAI consultants/experts need.	Teams are defensive or resistant in their engagements with RAI consultants/experts. For example, when teams receive RAI Specialists' input, they might push back against it and the recommended work. If the RAI work is done, it is done by RAI consultants or RAI experts, and not by the team. When teams provide RAI Specialists information, they might share too little/ too much information that makes it impossible for consultants/experts to understand the situation and help.	Teams are indifferent/halfhearted in their engagements with RAI consultants/experts. For example, when teams receive RAI consultants'/experts' input, they might be somewhat receptive to it but in a passive way – they want to be told exactly what to do. When teams provide RAI consultants/experts information, they might not consider what information is needed and wait to be told what information to share. This can lead to a lot of back-and-forth and wasted time.	Teams are open in their engagements with RAI consultants/experts. For example, when teams receive RAI consultants'/experts' input, they might embrace it and actively engage in learning about RAI and implementing recommendations. When teams provide RAI consultants/experts information, they might share adequate information and make available a point of contact who connects the RAI consultants/experts with the right people and resources (e.g., access to systems, telemetry data, RAI incidents reported).	Teams form true partnerships with RAI consultants/experts. For example, teams and RAI consultants/experts might work together to identify, understand, and address RAI problems. When teams provide RAI consultants/experts information, they might consider the audience– what background knowledge they have, and if unfamiliar with the work, what they need to know. As a result, teams brief RAI consultants/experts about the project at the appropriate level of detail. Teams keep RAI consultants/experts up to date on the latest developments, changes, or updates.

RAI Practice

What: *RAI Practice* includes the dimensions of RAI maturity that pertain to how teams perform specific RAI work. RAI Practices are often done in the service of high-level goals such as Fairness, Inclusiveness, and Reliability & Safety. We found that accomplishing high-level principles requires AI practitioners to perform the work of identifying, measuring, mitigating, and monitoring RAI risks.

Keep in mind that although the dimensions listed below are often the most visible forms of RAI work in organizations, they are shaped by the *Organizational Foundations* and *Team Approaches* dimensions. Maturity in RAI Practice can enable AI practitioners to influence and advance the RAI maturity of their team and organization.

Combining team assessments across a department can give a snapshot of the department's average maturity for each *RAI Practice* dimension (do not average across dimensions). Note that teams within a department can vary drastically in their maturity, thus the focus on efforts to increase maturity in RAI practice dimensions should be specifically tailored to teams, the type of their AI system, and its deployment domain.

Who: *RAI Practice* dimensions should be assessed by the organization's teams as the maturity of such dimensions reflects how they perform RAI work. Assessing the maturity on these dimensions can help practitioners and teams take stock of their ability to do the different kinds of sociotechnical work required to design, build, and maintain RAI systems.

Accountability

Accountability is the practice of taking responsibility for mitigating harms in AI products, features, and systems. An organization is accountable not only to the direct end-users of its products but also to its customers who use and deploy systems built by the organization.

The *Accountability* dimension should be assessed at an organizational level.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
The organization takes no responsibility for harm mitigation.	The organization takes minimal responsibility for harm mitigation. The organization offers customers and users superficial RAI documentation and disclaimers. The organization relies on end users to oversee system harms or only use the system in approved ways. The organization relies on customers who deploy the AI system to put human oversight mechanisms in place and do their own last-mile system testing without any additional guidance or instruction.	The organization takes some responsibility for harm mitigation. It recognizes the limits of human oversight, providing users and customers with detailed RAI documentation (e.g., Transparency Notes). The organization does not have RAI accountability policies in place. The organization may mitigate harms by using gating or system controls to limit and track deployments to prevent harmful uses. The organization identifies and documents the stakeholders that are responsible for troubleshooting, managing, operating, overseeing, and controlling the system during and after deployment.	The organization partners with users and customers in sharing responsibility for harm mitigation. The organization recognizes the limits of human oversight, providing guidance and training on human oversight considerations to users and customers to empower them to manage harms and failures. The organization has policies in place for RAI accountability. When possible, the organization builds harm mitigations into the system itself, designing and building systems to avoid failures.	The organization partners with users and customers in sharing responsibility for harm mitigation. It defines and documents methods used to evaluate whether oversight functions can be realistically accomplished by stakeholders, including the metrics used in the evaluations. When this is not possible, it provides guidance on evaluating oversight functions to the third party responsible for evaluating oversight functions. The organization has a documented plan for managing previously unknown failures as they surface, including feature and system rollback, processes for model updates, and notifying users and customers of system changes.

Transparency

Transparency is the practice of being open about the information necessary to understand AI systems. From an RAI perspective, being transparent isn't just about documenting and sharing technical information (e.g., datasets, models, and benchmarks) but also RAI-specific information such as known limitations, potential harms, and inappropriate uses of AI systems.

External transparency

External transparency is the practice of publicly sharing AI system information with users, customers, and stakeholders to clearly communicate the capabilities, limitations, and potential risks of AI systems. External transparency is accomplished using artifacts such as transparency notes, model or system cards, tutorials, online documentation, FAQs, and UI copy.

The *external transparency* dimension should be assessed at both the organizational level and team level.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
The organization does not share system information externally or only shares technical documentation. It may not be obvious that AI is used at all.	The organization shares technical system information (e.g., functionality, instructions for use) externally but does not include any information on RAI or sociotechnical considerations (e.g., explanations of system behaviors, known limitations, potential harms). For systems that have user interfaces, system information is not integrated into the user interface; the user must seek it out.	The organization shares both technical system information and RAI considerations, but the RAI content is boilerplate and lacks specificity (e.g., discusses RAI at a conceptual level, not specific to the model/system). For systems that have user interfaces, generic RAI system information is integrated into the user interface at a few key points.	The organization shares both technical system information and specific RAI considerations, like known limitations, potential harms, and responsible use of the model/system. For systems that have user interfaces, specific RAI system information is integrated into the user interface at a few key points. Specific RAI system information is also available via documentation, and possibly other formats (e.g.,	The organization shares both technical system information and specific RAI considerations with appropriate transparency. They adapt communications and documentation to correspond to the intended audience and evaluate the content and format for effectiveness. For systems that have user interfaces, system information is integrated throughout the user experience – e.g., readily available explanations of

	The information is difficult to find (e.g., it is only captured in stand-alone documentation, not accessible via other formats and requires searching to find).	Generic RAI system information is also available via documentation.	tutorials, notebooks, training videos) and is easy to access.	outputs are provided in easy-to understand language. Mechanisms are offered for any stakeholder to engage in a dialog or ask questions about this information. Transparency is seen as a continuous interaction and as systems evolve the information shared externally is updated.

Internal transparency

Internal transparency is the practice of systematically documenting information related to the design, development, and working of AI systems to ensure auditability, collaboration, reflections on potential RAI risks, and external transparency. Internal transparency is accomplished using artifacts such as code and system documentation, impact assessments, datasheets for datasets, system architecture diagrams, and model cards.

The *internal transparency* dimension should be assessed at both the organizational level and team level.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
The organization does not have any requirements or guidance for what information needs to be documented and how. Some information about datasets, models, decisions, processes, etc. might be captured in emails, meeting	The organization does not have any requirements or guidance for what information needs to be documented and how. Some information individuals deem important is documented at the team level, on an as-needed basis, but	The organization recommends teams document information about AI assets and offers templates (e.g., Impact Assessments, Datasheets for Datasets, Model Cards) but no or very little documentation is formally required.	The organization has clear requirements, templates, guidance, and supporting resources for documentation. Some standardized documentation exists across the organization but is not fully integrated into AI	The organization integrates required documentation into AI pipelines, workflows, and existing tools. Relevant information is documented consistently across the organization and kept up to date.

notes, presentations. Some information remains undocumented and dependent on individuals' memory. As a result, not all relevant information is captured or retrievable. When an individual leaves the project, information is lost.	this is done inconsistently throughout the organization.	Some standardized documentation exists but is created inconsistently across the organization, might be incomplete and difficult to find. Documentation is created towards the end of the AI development and deployment lifecycle.	pipelines, workflows, existing tools). Some documentation is created throughout the AI development and deployment lifecycle, but this is not yet consistent across the organization. Documentation is sufficient to facilitate auditing.	Documentation captures all necessary system design aspects (design & development choices, rationales, and assumptions) and RAI considerations (e.g., limitations, supported & unsupported uses, fairness assessments). Documentation is actively used to inform AI development and deployment decisions. It facilitates cross-team collaboration (e.g., other teams can use ML assets such as models, datasets, RAI mitigations) and auditing. The organization's documentation requirements, templates, practices and guidance are adopted by others in the industry and inform industry standards and policies.

Identifying, measuring, mitigating, and monitoring RAI risks

Identify – measure – mitigate is a recommended framework for addressing RAI risks. Monitoring completes the cycle. The following dimensions show what maturity entails for each step in this framework.

Identifying RAI risks

Identifying RAI risks is the practice of determining risks and issues specific to an AI system, its uses, stakeholders, and deployment contexts. This is usually the first step that AI practitioners take when understanding the potential harms arising from AI systems. These risks can be of different kinds such as fairness, reliability, safety, privacy, or security. The practice of identifying RAI risks includes activities such as doing impact assessments, stakeholder engagement, and red teaming.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
Teams lack awareness about what constitutes RAI risks. RAI risks or harms are identified and raised only by external parties (e.g., auditors, media, new employees).	Teams identify RAI risks in an ad-hoc manner, without conducting specific investigations (e.g., an employee stumbles upon it).	Teams conduct ad-hoc investigations to identify RAI risks (e.g., eyeballing data, speculating on risks, occasionally doing impact assessments). Teams engage in some reasoning about risks, but without involving affected stakeholders to understand how the risks affect them. However, teams might be unable to prioritize which risks to address and to manage trade-offs among AI system risks, benefits, and business requirements. This might result in paralysis – since a	Teams conduct systematic investigations to identify RAI risks using established practices such as an impact assessment. Teams' reasoning about risks is informed by stakeholder input. They have a clear awareness of risks, benefits and business requirements, but struggle to prioritize them and manage trade-offs.	Teams have a documented and structured plan to identify RAI risks (e.g., impact assessment), and this plan is integrated into the AI development and deployment lifecycle. Reasoning about risks is based on deep engagement with stakeholders, leveraging sociotechnical experts such as UX researchers. Teams engage in in-depth, data-driven cost-benefit analysis to make prioritization decisions and balance risks,

		perfect system is unachievable, they forgo RAI altogether (all or nothing thinking).		benefits and business requirements. Teams update and review the impact assessment at least annually, when new intended uses are added, and before advancing to a new release stage. (Goal A1.3 in the Microsoft RAI Standard v2)

Measuring RAI risks

Measuring RAI risks is the practice of assessing the extent or severity of RAI risks once they have been identified. Such assessment can be quantitative, qualitative, or a mix of both. Measurement helps to better understand different RAI risks, prioritize among different risks, and track progress in mitigating them. The practice of measuring RAI risks involves different forms of work such as data collection, using specific tools, collaborating with RAI experts and subject-matter experts, and defining performance metrics aligned with products or features. (See Appendix for methods and tools.)

Level 1	Level 2	Level 3	Level 4	Level 5
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
Teams do not conduct RAI-specific risk measurements, relying instead on general, mostly offline, aggregated AI performance metrics such as accuracy. Teams only consider stakeholders in relation to basic business needs (e.g., prioritizing business owners), and might not consider	Teams measure only some of the system's RAI-specific risks (e.g., quality of service harms), in an ad-hoc manner. Teams use existing data for measuring RAI risks. They might be aware of the dataset's limitations (e.g., fit for purpose and representativeness) but are unable to quantify them or	Teams measure multiple RAI-specific risks (e.g., known failures, quality of service harms, security threat modeling) and start acknowledging challenges with measuring different risks (e.g., difficulties quantifying representational harms). Teams sometimes collect additional data in an ad-hoc manner for measuring RAI risks. When possible, they try	Teams have a documented evaluation plan for different risk measurements, performance metrics, and error types that encompasses both quantitative and qualitative approaches to harm assessments. Teams collect additional data as part of a well-developed strategy for RAI risk measurement.	Teams conduct risk measurements on a regular basis, as part of system monitoring. Teams have an established process for continuously monitoring the system to refresh and retire data (e.g., does the data continue to be representative of new stakeholders, model changes, and new features).

stakeholders exposed to RAI risks.	address them by collecting more data. Teams measure predictable failures (e.g., false positives, false negatives) and analyze how they would impact stakeholders.	to quantify or address the dataset's limitations. Teams employ different approaches to risk measurement, going beyond predictable failures (e.g., aggregated vs. disaggregated evaluation, exploring use of RAI tools). Teams engage with RAI literature (e.g., research papers) to identify and measure how different factors might cause harms, but rarely work with RAI experts to validate their approaches.	Teams employ a wide range of RAI-specific risk measurement approaches (e.g., measuring across multiple populations with careful consideration of factors and groups) with an eye towards reproducibility of measures. Teams go beyond measuring risks and try to understand the root cause of problems (e.g., data representativeness vs. model limitations). Teams work with RAI experts and employ RAI best practices (e.g., using disaggregated evaluation or RAI tools such as RAI Dashboard and Toolbox), while simultaneously finding new ways to measure and track model performance (e.g., use-case specific error tolerances). Teams work with domain-specific subject matter experts (e.g., medical doctors if working on healthcare AI systems) to understand how different factors may impact system performance, including how such factors might vary across different groups.	Teams actively reflect on and improve their risk measurement practices (e.g., identifying new ways to surface and measure risks). Teams work not only with RAI and domain-specific experts, but also with members of identified demographic groups to understand the risks of and impacts associated with model behavior (e.g., performance disparities).

Mitigating RAI risks

Mitigating RAI risks is the practice of choosing the right strategy to address RAI risks once they have been identified and measured, both pre- and post-deployment. While some mitigations strategies are technical in nature (e.g., changing model parameters, collecting more data), others can take alternate forms such as UX/UI design interventions or RAI documentation. The practice of mitigating RAI risks involves different work such as understanding trade-offs between mitigation strategies, matching strategies to risks at different stages in the product lifecycle and creating a prioritization strategy to address different harms.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
Teams might apply some mitigation techniques without measurement of RAI risks, potentially causing more harm.	Teams use mitigation strategies guided by incomplete identification and measurement of risks. For example, they might aim for increased data representativeness to mitigate fairness harms, but without careful consideration of specific affected groups and the specific harms they experience.	Teams' choice of mitigation strategy is guided by a more complete identification and measurement of risks. However, Teams are not yet able to optimally match mitigation strategies to identified risks or to apply the mitigation at the optimal stage in the ML development and deployment lifecycle. Therefore, the chosen mitigation strategy might not be the most effective, but it does not lead to further harms.	Teams' choice of mitigations strategy is guided by a thorough understanding (identification and measurement) of risks and of trade-offs among various mitigation strategies. Teams are able to optimally match mitigation strategies to identified risks. Focus for mitigation is limited to only 1-2 stages in the ML development and deployment lifecycle. **And/or:** Teams do not have a monitoring system in place that enables them to identify risks on an ongoing basis.	Teams' choice of mitigations strategy is guided by a thorough understanding (identification and measurement) of risks and of trade-offs among various mitigation strategies. Teams are able to optimally match mitigation strategies to risks and consider mitigations at every stage in the ML development and deployment lifecycle. Understanding that it is not possible to identify and mitigate all risks for all groups, teams have a monitoring system in place and a prioritization strategy that they document and adjust periodically.

Monitoring RAI risks

Monitoring RAI risks is the practice of keeping track of AI systems and features post deployment to check for RAI risks such as those of fairness, reliability and safety, security, and human-AI interaction and collaboration. While monitoring AI systems for performance guarantees such as latency is a common practice, monitoring AI systems for RAI risks requires AI practitioners to take on different things such as outlining their safety posture, checking for overreliance on AI, employing extensible telemetry frameworks, and doing continuous adversarial testing.

Level 1	Level 2	Level 3	Level 4	Level 5
				
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
Teams do not actively perform system monitoring. They rely, for instance, on external signals such as user reports, but those reports might not be systematically reviewed.	Teams take an ad-hoc approach to system monitoring (e.g., if a customer points out an issue, they will sometimes do a manual investigation). Teams conduct some offline testing for RAI risks (e.g., manual probing and model testing). Teams do system monitoring mainly to check for performance guarantees (e.g., latency, service availability) and ensure that performance SLAs are met (e.g., latency < 300ms).	Teams take a proactive approach to system monitoring (e.g., considering aspects of telemetry design early in the development pipeline). They also often check social media and systematically review user reports as part of system monitoring. Teams start having discussions about specific RAI risks (e.g., what reliability and safety mean in their context, what is their security posture). Teams do system monitoring mainly to ensure performance guarantees, and to a lesser extent to check for some downstream harms (e.g., evidence of overreliance, model degradation, out of context model usage, and fairness-related harms).	Teams take a systematic approach to system monitoring (e.g., use of telemetry frameworks, clear ways to incorporate learnings from user feedback and telemetry into system design, and established processes for failure awareness). Teams do systematic analysis of RAI risks (e.g., fairness-related harms, data poisoning, overreliance), have clearly defined measurement goals (e.g., as part of safety specs), and have ways to show that RAI goals were achieved. Teams do system monitoring to not only ensure that the system does not cause downstream harm, but also keep an eye out for unforeseen failures and issues (e.g., backward compatibility,	Teams take a systematic and scalable approach to system monitoring. They have pre-defined plans for failures (e.g., shut down or replace ML components), employ extensible telemetry frameworks (new measures can be added easily and quickly), and try to find new ways of dealing with AI-specific monitoring (e.g., how to identify and measure issues when dealing with eyes-off telemetry). Teams analyze the RAI threat surface and define the safety spec in a cross-disciplinary manner (instead of just data scientists or engineers doing the analysis, the team works together to establish what safety means for different parties (e.g., business vs. customer vs. user).

			unintended uses, black-box attack vectors).	Teams do continuous internal testing as part of system monitoring (e.g., adversarial testing, red teaming, bug bashing) to proactively identify and mitigate RAI risks.

AI privacy and security

AI poses unique challenges for privacy and security. Mature RAI practice takes into consideration and addresses them, instead of relying on traditional approaches.

AI privacy

AI privacy is the practice of identifying, measuring, and mitigating privacy risks specific to AI products, features, and systems. This dimension complements existing privacy frameworks (e.g., Microsoft's Privacy Standard) by adding AI-specific considerations such as member identification through model leakage, training data memorization, and federated learning.

While there are different maturity models for privacy, there aren't specific maturity models for *it* from an RAI perspective. The dimensions should thus be seen as complementary to, and not a replacement for, already-existing privacy maturity models. [For references to related maturity models see the appendix.] It is possible for an organization to be highly mature in traditional privacy, but less mature in *AI privacy*. However, high levels of *AI privacy* maturity require high levels of traditional privacy maturity.

The *AI privacy* dimension should be assessed at both the organizational level and team level.

Level 1	Level 2	Level 3	Level 4	Level 5
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
Teams and privacy personnel (e.g., privacy managers) have a good understanding of general privacy risks; However, they are not aware of privacy risks specific to AI systems. The organization's privacy processes (e.g., privacy reviews, threat modeling, incident response) address general legal privacy legal	Teams and privacy personnel are aware of AI-specific privacy risks such as member identification through model leakage, memorization of training data, reverse prompt engineering issues. However, they do not have the knowledge, guidance, processes, or resources to	Teams and privacy personnel are aware of AI-specific privacy risks and employ some AI-specific mitigations (e.g., differential privacy, federated learning, or use of synthetic data). However, in the absence of clear guidance, the choice of mitigation might not be optimal. For example, teams might use differential privacy	Teams and privacy personnel have knowledge of a wide range of means to identify, measure and mitigate AI-specific privacy risks. To address AI privacy risks, they engage in critical discussions about ensuring a match between privacy goals, risks, and mitigations.	The organization builds on knowledge gained from previous AI projects to create guidance and reusable practices to identify, measure, and mitigate AI privacy risks. AI privacy is integrated into the organization's privacy processes.

requirements but do not account for AI-specific privacy risks.	identify, measure and mitigate AI-specific privacy risks.	because they are aware it is beneficial, but they lack evidence that it is effectively addressing the AI system's specific privacy risks. Pockets of AI privacy expertise begin to emerge in the organization, including among privacy personnel, but AI privacy is handled on a case-by-case basis. Addressing difficult AI privacy scenarios depends on being able to identify and leverage the few experts in the organization.	The organization is beginning to develop policies, guidance, and requirements specific to AI privacy. Even though formal processes and guidance might not exist, AI practitioners and privacy personnel document their AI privacy decisions, enabling auditability.	Privacy experts in the organization work to advance the state of the art of AI privacy, through research, guidance, and tooling. The organization's privacy personnel is proactive in anticipating regulatory changes and consults external institutions and governing bodies on the development of AI privacy regulations.

AI security

AI security is the practice of identifying, measuring, and mitigating security risks specific to AI products, features, and systems. This dimension complements existing security frameworks by adding AI-specific considerations such as model evasion, adversarial attacks, and other aspects captured in AI-specific security frameworks (e.g., MITRE ATLAS).

While there are different maturity models for security, they are not from a RAI perspective. This dimension should thus be seen as complementary to, and not a replacement for, already-existing security maturity models (see the Appendix for such models). It is possible for an organization to be highly mature in traditional security, but less mature in *AI security*. However, high levels of *AI security* maturity require high levels of traditional security maturity.

Level 1	Level 2	Level 3	Level 4	Level 5
Latent	**Emerging**	**Developing**	**Realizing**	**Leading**
Teams have a good understanding of general security risks; However, they	Teams are concerned about AI-specific security issues but apply traditional	AI security risks (e.g., data poisoning) are not assumed to already be entirely covered by	Teams might have some processes in place for addressing AI-specific security	Comprehensive adversarial testing and threat modeling of AI systems is integrated into

are not aware of security risks specific to AI systems.	methodologies to AI systems such as those described in the Security Development Lifecycle (SDL)	existing security processes such as those in the SDL. For example, teams are aware of some possible adversarial risks to AI systems (e.g., model evasion), and that these adversarial threats require specific mitigations. Teams may take ad-hoc steps to address AI security risks (e.g., red teaming to evaluate the effectiveness of AI-specific security mitigations).	issues. For example: The SDL might include looking at adversarial risks as part of system development; Incident response process for the organization has been updated to include adversarial attacks. Teams start measuring and monitoring their security posture (e.g., how many services run AI models, what is the provenance of these models, how many of them have gone through threat modeling that includes AI adversarial threats). Teams know about and refer to AI-specific security frameworks such as MITRE ATLAS for threat modeling at different stages (e.g., model access vs. execution, data curation vs. training).	the AI development pipeline – done on a regular basis (e.g., when a substantial change is made to the model), at scale, and using automated tools (e.g., Counterfit). Teams understand the evolving nature of AI security threats, and can detect, reflect, and respond to new issues in a reasonably short time. The organization's security guidance, processes, frameworks, tooling are adopted by others in the industry and inform industry standards and policies.

References

Becker, J., Knackstedt, R., & Pöppelbuß, J. (2009). Developing maturity models for IT management: A procedure model and its application. *Business & Information Systems Engineering, 1*, 213-222.

De Bruin, T., Rosemann, M., Freeze, R., & Kaulkarni, U. (2005). Understanding the main phases of developing a maturity assessment model. In *Australasian Conference on Information Systems (ACIS)* (pp. 8-19). Australasian Chapter of the Association for Information Systems.

Hornick, M. (2020). *A Data Science Maturity Model for Enterprise* [White paper]. Oracle. https://blogs.oracle.com/machinelearning/post/a-data-science-maturity-model-for-enterprise-assessment

Mettler, T. (2011). Maturity assessment models: a design science research approach. *International Journal of Society Systems Science, 3*(1-2), 81-98.

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

Pringle, T. & Zollen, E. (2018). *How to Achieve AI Maturity and Why It Matters* [White paper]. Ovum. ai-maturity-model-whitepaper.pdf (amdocs.com)

Sanderson, C., Douglas, D., Lu, Q., Schleiger, E., Whittle, J., Lacey, J., Newnham, G., Hajkowicz, S., Robinson, C., & Hansen, D. (2023). AI Ethics Principles in Practice: Perspectives of Designers and Developers (arXiv:2112.07467). arXiv. http://arxiv.org/abs/2112.07467

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2020). Principles to Practices for Responsible AI: Closing the Gap (arXiv:2006.04707). arXiv. http://arxiv.org/abs/2006.04707

Sheng, A., Wu, E., Sarkar, E., Pratt, A., Kudrle, T., Sullivan, R., & Iansiti, M. (2020). *Rethinking the enterprise* [White paper]. Keystone.ai. https://fs.hubspotusercontent00.net/hubfs/6724850/Keystone-Strategy-Rethinking-the-Enterprise-White-Paper-May-2021.pdf

Shneiderman, B. (2020). Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems, 10*(4), 1–31. https://doi.org/10.1145/3419764

Vaish, R., Agrawal, A., Kapoor, S., & Parkin, R. (2021). *AI Maturity Framework Enterprise Applications* [White paper]. IBM. https://www.ibm.com/downloads/cas/OB8M18WR

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. *CEUR Workshop Proceedings*, *2808*. https://ceur-ws.org/Vol-2808/Paper_16.pdf

Appendix

Responsible AI Tools

Microsoft collection of RAI resources

Microsoft's RAI Standard v2

Microsoft RAI Toolbox

HAX Toolkit

Aether Data Documentation Template

Fairness Checklist

AI Security Guidance

AI/ML Maturity Models

White papers

Accenture. (2021). Art of AI Maturity

Boston Consulting Group (2021) Responsible AI Maturity

Element AI (2020). The AI Maturity Framework

IBM (2021). AI Maturity Framework Enterprise Applications

Keystone.ai (2021) Rethinking the Enterprise

Microsoft (2018). AI Maturity & Organizations

NIST (2023) AI Risk Management Framework

Oracle (2020). A Data Science Maturity Model for Enterprise Assessment

Ovum (2018). How to Achieve AI Maturity and Why It Matters

Salesforce (2021). Ethical AI Maturity Model

Research papers

Akkiraju, R., Sinha, V., Xu, A., Mahmud, J., Gundecha, P., Liu, Z., Liu, X., & Schumacher, J. (2020). Characterizing Machine Learning Processes: A Maturity Framework. In D. Fahland, C. Ghidini, J. Becker, & M. Dumas (Eds.), *Business Process Management* (pp. 17–31). Springer International Publishing. https://doi.org/10.1007/978-3-030-58666-9_2

Alsheibani, S., Cheung, Y., & Messom, C. (2019) Towards an Artificial Intelligence Maturity Model: From Science Fiction To Business Facts. *PACIS*, p. 46. 2019. https://aisel.aisnet.org/pacis2019/46

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software Engineering for Machine Learning: A Case Study. *Proceedings - 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice, ICSE-SEIP 2019*, 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

Krijger, J., Thuis, T., de Ruiter, M., Ligthart, E., & Broekman, I. (2022). The AI ethics maturity model: A holistic approach to advancing ethical data science in organizations. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00228-7

Nick, G., Kő, A., Szaller, Á., Zeleny, K., Kádár, B., & Kovács, T. (2022). Extension of the CCMS 2.0 maturity model towards Artificial Intelligence. *IFAC-PapersOnLine*, *55*(10), 293–298. https://doi.org/10.1016/j.ifacol.2022.09.403

Privacy and Security Maturity Models

White papers

AICPA (2020) The Privacy Management Framework

Association of Corporate Counsel (2019) U.S. States' Privacy Laws Capability Maturity Model

MITRE (2019) Privacy Maturity Model

NIST (2018) Cybersecurity Framework

NIST (2020) Privacy Framework

NIST (2022) Cybersecurity Capability Maturity Model

SANS Security Awareness Security Awareness Maturity Model

Research papers

Almuhammadi, S., & Alsaleh, M. (2017). Information security maturity model for NIST cyber security framework. *Computer Science & Information Technology (CS & IT)*, *7*(3), 51-62. https://doi.org/10.5121/csit.2017.70305

Rabii, A., Assoul, S., Ouazzani Touhami, K. and Roudies, O. (2020), "Information and cyber security maturity models: a systematic literature review", *Information and Computer Security*, Vol. 28 No. 4, pp. 627-644. https://doi.org/10.1108/ICS-03-2019-0039

Related Papers

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

Buhmann, A., & Fieseler, C. (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society*, *64*, 101475. https://doi.org/10.1016/j.techsoc.2020.101475

de Laat, P. B. (2021). Companies Committed to Responsible AI: From Principles towards Implementation and Regulation? *Philosophy & Technology*, *34*(4), 1135–1193. https://doi.org/10.1007/s13347-021-00474-3

Dove, G., Halskov, K., Forlizzi, J., & Zimmerman, J. (2017). UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 278–288. https://doi.org/10.1145/3025453.3025739

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3290605.3300830

Ibáñez, J. C., & Olmeda, M. V. (2022). Operationalising AI ethics: How are companies bridging the gap between practice and principles? An exploratory study. *AI & SOCIETY*, *37*(4), 1663–1687. https://doi.org/10.1007/s00146-021-01267-0

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399.

Kelley, S. (2022). Employee Perceptions of the Effective Adoption of AI Principles. *Journal of Business Ethics*, *178*(4), 871–893. https://doi.org/10.1007/s10551-022-05051-y

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376445

Mao, Y., Wang, D., Muller, M., Varshney, K. R., Baldini, I., Dugan, C., & Mojsilovic, A. (2019). How data scientists work together with domain experts in scientific collaborations: To find the right answer or to ask the right question? *Proceedings of the ACM on Human-Computer Interaction*, *3*(GROUP), 1–23. https://doi.org/10.1145/3361118

Mettler, T. (2010). Thinking in Terms of Design Decisions When Developing Maturity Models: *International Journal of Strategic Decision Sciences*, *1*(4), 76–87. https://doi.org/10.4018/jsds.2010100105

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds and Machines*, *31*(2), 239–256. https://doi.org/10.1007/s11023-021-09563-w

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

Orr, W., & Davis, J. L. (2020). Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society*, *23*(5), 719–735. https://doi.org/10.1080/1369118X.2020.1713842

Passi, S., & Jackson, S. J. (2018). Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–28. https://doi.org/10.1145/3274405

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33-44).

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

Ryan, M., Christodoulou, E., Antoniou, J., & Iordanou, K. (2022). An AI ethics 'David and Goliath': Value conflicts between large tech companies and their employees. *AI & SOCIETY*. https://doi.org/10.1007/s00146-022-01430-1