Issue Brief

A Matrix for Selecting Responsible AI Frameworks

Authors
Mina Narayanan
Christian Schoeberl

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

June 2023

Executive Summary

Organizations have a growing number of tools at their disposal to implement responsible AI systems, or systems that minimize unwanted risks and create beneficial outcomes. However, it is not always clear how to select and apply these tools. This paper provides a way for organizations to systematically characterize one type of tool—namely, process-based frameworks—that accommodates their specific needs. Process frameworks for AI provide a blueprint to ensure that organizations are prepared to meet the challenges and reap the benefits of AI systems. They can help an organization prioritize aspects of system design, build lines of accountability into product development teams, and engage with impacted communities, among many other critical functions. Without an action plan to follow, organizations would struggle to establish the infrastructure, resources, and capabilities needed for responsible AI.

However, process frameworks vary in their level of specificity, with many erring on the side of generality to accommodate flexibility in implementation. Although this can be desirable in certain circumstances, it can be burdensome for organizations that want to choose a framework but lack experience in implementing responsible AI. Devising a standard way of comparing a large number of frameworks takes time and energy that organizations may not have. First, organizations may struggle to determine who can use a framework. While some frameworks name a target audience, many do not. Even when an audience is mentioned, they are commonly described in general terms, which makes it difficult to assign responsibilities for operationalizing a framework. Second, once an audience is identified, it may still be difficult to discern which needs are satisfied by that framework.

In this brief, we reviewed and organized more than 40 existing responsible AI frameworks put forth by a variety of companies, international organizations, government bodies, and non-governmental organizations and mapped them onto a matrix that can help organizations better understand, select, and implement responsible AI in a way that best fits their needs. The matrix is focused on the user, appealing to the people who are most directly involved in implementing frameworks within an organization that builds or uses AI, namely, the Development and Production team and the Governance team. To help these users select the frameworks that will best serve their needs, the matrix adds a Utility dimension, further classifying frameworks according to their respective areas of focus: an AI system's components, an AI system's lifecycle stages, or characteristics related to an AI system. The matrix provides a structured way of thinking about who could benefit from a framework and how that framework could be used, which helps organizations precisely apply frameworks and understand the utility of each framework relative to guidance that already exists.

Table of Contents

Background

Momentum has grown within government and industry to organize tools for implementing responsible AI systems, or systems that minimize unwanted risks and produce beneficial outcomes. The National Institute of Standards and Technology (NIST) has expressed intent to build a Trustworthy and Responsible AI Resource Center that will house these tools in a centralized location online.[1] The Organisation for Economic Co-operation and Development (OECD) has created a taxonomy for comparing technical, procedural, and educational tools and practices for implementing responsible AI.[2] Even standalone GitHub repositories have assembled resources for implementing responsible AI systems.[3] However, none of these initiatives currently expands on the utility of a responsible AI tool that has gained widespread recognition: process frameworks.

Process frameworks are qualitative guidance that enable systematic and actionable tasks for implementing responsible AI systems. Implementation is used here to mean any action that brings an AI system into existence or supports its use, such as maintaining the system or setting up channels for users to seek recourse from harm inflicted by the system. Detailed descriptions of process frameworks can be found in Appendix B. The proliferation of these frameworks suggests that organizations have more resources at their disposal but does not imply that organizations know how to properly leverage them. This paper presents a system of categorizing process frameworks that enables organizations to select frameworks and capitalize on their strengths while avoiding their weaknesses.

Process Frameworks

Process frameworks boast a number of advantages. They are flexible and can be modified to fit an organization's unique circumstances, which also broadens their appeal. This contrasts with technical frameworks that can only be applied to AI systems that meet specific design requirements. A small team developing an image recognition system, for example, will likely operationalize frameworks differently than an organization with a dedicated division for building test beds for large language models. Process frameworks can account for the different needs of these organizations by steering away from rigid requirements, and they can complement technical tools by contextualizing quantitative outputs. For example, they can structure documentation for models and data, outline risk management frameworks, articulate the steps involved in an impact assessment or audit, and list actions to prevent failure modes. They can act as a process shell into which teams can plug in the appropriate technical tools or provide scaffolding for more granular standards.[4]

While generality can be a strength of process frameworks, it poses a challenge when determining who should use these frameworks. Some frameworks explicitly mention an audience, but many do not. Even when a target audience is stated, it sometimes consists of such a broad swath of people, such as "AI practitioners" or "groups responsible for implementing AI systems," to not be immediately useful in practice.[5] An ambiguously defined audience can prevent people who actively participate or lead the selection of a process-based framework from assigning responsibility to the most qualified stakeholders. Although this lack of specificity may pose only a minor challenge for some organizations, those with more limited resources may be discouraged from using the framework altogether.

Once a framework is in the hands of an individual or a team with the appropriate responsibilities, figuring out how to best leverage the framework can still be challenging. The needs that a framework meets may not be immediately apparent. Without additional support to guide selection, users may not choose the optimal framework for implementation.

For this project, we reviewed approximately 40 existing responsible AI frameworks and mapped them onto a matrix that can help organizations select and use AI frameworks based on their specific requirements and needs. The matrix consists of two dimensions: a User dimension and a Utility dimension. Users are people responsible for implementing frameworks within an organization that builds or uses AI systems. We split this dimension into two categories—the Development and Production team and the Governance team—as these are the teams most often in charge of designing, developing, and productizing AI systems or implementing policy for their use. After the user is identified and has articulated the specific needs of the team and the organization writ large, the Utility dimension can assist with further targeting of relevant frameworks based on the framework’s respective area of focus: components, lifecycle, or characteristics.

The 40 free and openly available process frameworks that were mapped onto the matrix move beyond defining principles by outlining steps for implementing responsible AI within an organization. More details about how frameworks were selected can be found in Appendix A. The populated matrix helps organizations filter through a large number of frameworks and serves as a launchpad for organizations to explore other frameworks.

Matrix for Responsible AI

Tables 1 and 2 contain the matrix. The tables display the names of frameworks that fall into several combinations of categories from the User and Utility dimensions. The gray cells in the tables indicate that no frameworks were found for that particular combination of categories. An organization looking for frameworks that fall into a gray cell can try blending other frameworks in the matrix together to produce a hybrid tool that achieves a similar effect. An organization can also seek out new frameworks, or create their own.

Table 1: First Part of Matrix for Implementing Responsible AI

	Characteristics	Components	Lifecycle
Development and Production Team	- Microsoft Responsible AI Standard, v2[6] - Designing Trustworthy AI[7] - AI Ethics Framework for the Intelligence Community[8] - Google's Responsible AI practices[9] - Hazard Contribution Modes of ML Components[10]	- Data Statements for NLP[11] - Data Readiness Report[12] - Model Cards[13] - Model Info Sheets[14]	
Governance Team	- Salesforce's AI Ethics Maturity Model[15] - ECP's AI Impact Assessment[16] - Rolls Royce's The Aletheia Framework 2.0[17] - Explaining decisions made with AI[18] - WEF's AI Oversight Toolkit for Boards of Directors[19] - Ethics Guidelines for Trustworthy AI[20] - TAII Framework for Trustworthy AI Systems[21] - AI Ethics Impact Group[22] - Machine Intelligence Garage's Ethics Framework[23]		- DIU's Responsible AI Guidelines[24] - Cognitive Project Management for AI[25] - Hard Choices in AI[26] - SMACTR Internal Algorithmic Auditing Framework[27] - NIST AI Risk Management Framework[28] - Reviewable Automated Decision-Making[29] - BSA's Framework to Build Trust in AI[30]

Table 1 shows frameworks that fall under the two categories of the User dimension and the individual categories from the Utility dimension. Many of the framework names are abbreviated because of space limitations.

Table 2: Second Part of Matrix for Implementing Responsible AI

	Components & Lifecycle	Components & Characteristics	Lifecycle & Characteristics	Components, Lifecycle, & Characteristics
Development and Production Team	- Datasheets for Datasets[31] - Test and Evaluation Framework for Multi-Agent Systems of Autonomous Intelligent Agents[32] - Towards Accountability for Machine Learning Datasets[33]	- ATARC's ML Model Transparency Assessment[34] - Responsible bots: 10 guidelines for developers of conversational AI[35]		
Governance Team	- Partnership on AI's ABOUT ML Reference Document[36] - capAI[37] - AI Fairness Checklist[38] - National Fair Housing Alliance's Purpose, Process, and Monitoring Framework[39] - Reward Reports for Reinforcement Learning[40]	- Model AI Governance Framework[41] - FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity[42] - Guidance on the AI Auditing Framework[43] - Towards a Standard for Identifying and Managing Bias in AI[44]	- What to Do When AI Fails[45] - Guidance on the Ethical Development and Use of AI[46]	- GAO's Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities[47] - System Cards for AI-Based Automated Decision Systems[48] - OECD Framework for the Classification of AI Systems[49] - DOE AI Risk Management Playbook[50]

Table 2 shows frameworks that fall under the two categories of the User dimension and interactions between categories from the Utility dimension.

User Dimension

Users are people responsible for implementing frameworks within an organization that builds or uses AI systems. They are usually not the end users of an AI system, but they could be. We identify two types of Users: Development and Production teams and Governance teams. Development and Production teams typically include engineers, product managers, data scientists, domain experts, and user researchers. As such, the frameworks most suited for these roles tend to focus on processes that inform or are integral to AI system design, engineering, operation, maintenance, or monitoring. Members of the Development and Production teams who are charged with implementing a particular framework may be tasked with understanding a model's sensitivity to different inputs or accounting for relevant demographic information when annotating data.

Governance teams usually include individuals serving in executive roles or the owners of the organization. The frameworks most suited to Governance teams then deal with functions such as oversight or management. Following the framework's guidance, Governance teams may be tasked with engaging the public for comment on AI systems, managing third-party auditor relationships, or evaluating the impact of AI systems on supply chains.

Frameworks Suited for Development and Production Teams

A framework that could be useful for Development and Production teams includes AI system design, engineering, operation, maintenance, and monitoring processes. These processes may include understanding an AI model's sensitivity to different inputs or minimizing personal data in the model's training stage through perturbation or federated learning. These and other related processes are typically implemented by people who have close proximity to the system in the development or production stages, including engineers, developers, data scientists, testers, and operators. For example, the framework *Test and Evaluation Framework for Multi-Agent Systems of Autonomous Intelligent Agents* suggests conducting black-box testing methods such as equivalence partitioning, boundary value analysis, state transition testing, and combinatorial testing for commercial off-the-shelf technologies, which are processes that are likely geared towards engineers.[51]

Frameworks suited for Development and Production teams may also describe processes that inform the design, engineering, operation, maintenance, and monitoring of AI. Domain experts who raise technical, legal, or social considerations that engineers may overlook are well-equipped to implement these processes. These are experts who may be familiar with the operating environment or exploitable parts of the system. Alternatively, they may be social science experts such as user researchers or human factors experts who understand how people will perceive or interact with the AI system. They can advise on ways of disclosing the limitations of an AI system to its users or defending against adversarial attacks, for example. Therefore, people on Development and Production teams who have deep knowledge about the inner workings of the system, its operation, or the context in which it will be applied are best positioned to effectively operationalize these frameworks.

The framework *Hazard Contribution Modes of Machine Learning Components*, for instance, calls for model validation techniques that are comprehensive and contextually relevant.[52] While developers should be involved in model validation, they can benefit from the knowledge of domain experts who can determine whether requirements have been met in a way that does not compromise other factors important to the model's functioning. *Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development,* another framework that is suited for Development and Production teams, emphasizes communicating system degradation to stakeholders.[53] While engineers can create the digital communication pathways that facilitate this communication, user experience designers can customize these pathways to maximize their effectiveness.

Frameworks Suited for Governance Teams

Frameworks that can be useful for Governance teams contain processes that are needed to perform oversight, management, or compliance functions for AI systems or the personnel working on these systems. These processes typically relate to evaluating the impact or ensuring the sustainability of an organization that develops or uses AI systems. They are usually overseen by strategic or management-level professionals such as executives, business owners, or board members. Frameworks suited for Governance teams often include requirements and recommendations for engaging with the public for comment on AI systems, providing due process mechanisms, contracting with independent risk or auditing agencies, or assessing the economic impact of AI systems on critical functions. The *OECD Framework for the Classification of AI Systems*, for example, recommends an impact assessment of AI on human rights and democratic values.[54] *Guidance on the Ethical Development and Use of Artificial Intelligence*, another framework relevant for Governance teams, encourages training and awareness raising for personnel to ensure they have the appropriate knowledge to work with AI systems.[55] Both of these activities can be organized by Governance professionals.

Most of the frameworks classified as suited for Governance teams in the matrix also contain processes for ensuring responsible AI that correspond to the responsibilities and functions of Development and Production teams. The frameworks classified as suited for Development and Production teams, on the other hand, exclusively contain processes directly relevant to Development or Production professionals.

Some processes, like those related to handling sensitive client data, could reasonably appear in frameworks relevant for both Development and Production teams and Governance teams. At the same time, there may be individuals or functions within a given organization whose responsibilities may not exclusively fall within the realm of Development and Production teams, or that of Governance teams; these can include legal, ethical, compliance (such as data protection officers or data export controllers), records management, civil liberties, and privacy professionals. Individuals and teams working in these professions can still find responsible AI frameworks that suit their respective functionalities. *Guidance on the AI Auditing Framework,* for example, reiterates how data protection officers should be knowledgeable about data protection legislation as well as the nature and use of the data itself.[56]

Utility Dimension

Once the most appropriate users of a framework are identified, the Utility dimension can assist with identifying frameworks that meet their specific needs. This dimension organizes frameworks into three categories — Components, Lifecycle, and Characteristics — where each category satisfies different needs. The three columns of Tables 1 and 2 contain frameworks that fall under one or more categories of the Utility dimension. The categories are not mutually exclusive, so a framework could belong to several simultaneously, as Figure B1 in Appendix B demonstrates. Appendix B also illustrates how frameworks are organized according to the Utility dimension in more detail. Table 3 highlights when each Utility category is more or less useful.

Table 3: Utility Dimension Descriptions and Example Uses

	Description	When Useful	When Less Useful
Components	Framework is focused on an AI system's components, such as data or models.	For considering the capabilities, impacts, benefits, and risks of an AI system's constituent parts	Parts of an AI system are abstracted away.
Lifecycle	Framework is focused on stages of an AI system's lifecycle.	- For structuring tasks throughout an AI system's lifecycle - For identifying resource needs at different milestones - For pinpointing when risk arises and who should manage it	Actions are not sensitive to time or tied to a stage of the AI system's lifecycle.
Characteristics	Framework is organized around one or more characteristics, such as explainability or privacy.	- For connecting AI products to desired business and societal outcomes - For monitoring progress on organizational goals	Organizations are not invested in operationalizing characteristics.

Components

Components frameworks disaggregate an AI system into smaller pieces and lay the groundwork for analysis of these components. Some frameworks scrutinize the algorithms and training, validation, and test data of a system, while many are dedicated only to data. Data has likely received a great deal of attention from the reviewed frameworks because it is a critical input from which an AI model learns about the world. *Data Readiness Report* is one such framework that zeroes in on data, proposing iterative documentation of data profiles and remediations to the data.[57]

People who need to consider the capabilities, impacts, benefits, and risks of an AI system's constituent parts, such as concerning interactions between a system's components, can use these frameworks. Frameworks that decompose an AI system into parts raise considerations about the data and algorithmic methods that an AI model employs. Characteristics about the data, such as its sparsity and provenance, directly shape a model's ability to faithfully represent concepts of interest. Different types of data, such as image or tabular data, are suited to different use cases and may require specific algorithms for processing. The algorithm itself, and the way it was tuned to a particular setting, will influence how an AI system performs in operational environments where the stakes may be high. Components frameworks can also focus on parts of an AI system that aggregate models and data together, such as chatbots or AI-enabled services.

Components frameworks are least helpful in situations where an AI system and its impacts are considered holistically to answer questions about the utility or value of the system. An executive evaluating an AI system's performance in the market or a manager tracking improvements in employee efficiency after the introduction of an AI system need to know less about a system's components and more about the sum of its parts.

Lifecycle

The lifecycle of an AI system consists of several stages that are interdependent and sometimes cyclical. Frameworks assign different names to the lifecycle stages, but they often include some combination of design, data collection and processing, training, evaluation, deployment, monitoring, and decommissioning. However, there are exceptions: the *Cognitive Project Management for AI Methodology* introduces a business understanding phase to consider whether the problem at hand should be solved by AI at all, and if so, the criteria for project success.[58] Additionally, the *NIST AI Risk Management Framework* (AI RMF) is organized around functions that are not necessarily tied to a specific lifecycle stage. Nevertheless, the AI RMF is still categorized as a Lifecycle framework because the functions can be implemented throughout various stages of an AI system's lifecycle. The Lifecycle category is important for understanding how an AI system moves from ideation to a fully deployed product, and how different stages influence each other.

Stakeholders who structure tasks throughout an AI system's lifecycle and identify resource needs at different milestones can benefit from Lifecycle frameworks because they frame the AI system lifecycle as stages with dependencies. The selection of success metrics in the project conception stages will inevitably influence the types of tests that the system is subject to later in its lifecycle. Identifying these dependencies, as well as differences in domain knowledge, resources, and artifacts needed to complete each stage, can ensure that the stages support rather than undermine one another. Frameworks that belong to the Lifecycle category aid in pinpointing when risk can arise and who should be involved in managing it. If the AI system performs unexpectedly while in development, then the data scientist or engineer building the system and the project manager responsible for structuring development activities should be involved in addressing the problem. On the other hand, if the system acts in unintended ways after deployment, then the product owner responsible for the performance of the system should also be consulted.

The Lifecycle category can structure thinking around an AI system's evolution, but it is less useful when actions are not sensitive to time or tied to a stage of the AI system's lifecycle. Creating a culture of safety and ensuring that an organization fosters diverse perspectives are examples of actions that persist without dependence on a lifecycle stage. Tables 1 and 2 in Appendix C depict how the stages of several frameworks overlap.

Characteristics

AI systems are often characterized according to desirable or undesirable properties. Desirable characteristics such as safety, robustness, or resilience represent the values that systems and the people building or using these systems should strive toward. Sometimes, attention is drawn to undesirable properties of AI systems, such as imbalanced data or model instability, in an effort to single out and avoid hazardous states.[59] Good or bad, these characteristics are defined by social and cultural norms that change over time.

Many process frameworks group procedures or actions together, according to characteristics. Guidance from the United Kingdom's Information Commissioner's Office and The Alan Turing Institute, for example, helps organizations enhance the explainability of their AI processes, services, and decisions.[60] Stakeholders that wish to tie activities to desired business and societal outcomes can derive value from Characteristics frameworks. These frameworks can help stakeholders think through the rationale of their actions and monitor progress on organizational goals, which serve important normative and management functions.

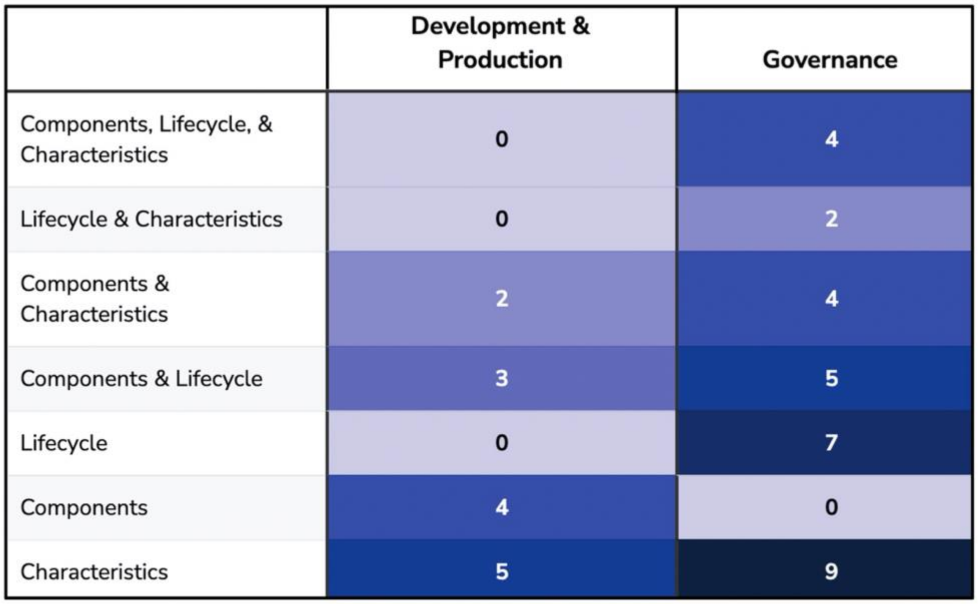
While a framework centered on characteristics can guide people's actions toward better outcomes, it can invite harmful practices such as ethics washing. Organizations that refuse to operationalize values or insincerely portray values as priorities render Characteristics frameworks ineffective. Table C3 in Appendix C shows how several frameworks focus on overlapping characteristics.

Tying the User and Utility Dimensions Together

The User and Utility dimensions are most effective when used together. With sufficient organizational buy-in and resources, the matrix can encourage tracking of team member involvement for different parts of an AI system from its inception to retirement, as well as delimiting the goals that team member actions serve. Figure 1 is a heatmap showing the number of frameworks that fall into the combinations of categories from the User and Utility dimensions. The relatively high number of Characteristics frameworks can likely be explained by the prominent role that desirable characteristics such as accountability and robustness have in shaping discussions about responsible AI.

Frameworks that fall under the Components and Lifecycle categories are oriented towards different users. Frameworks in the matrix that focus exclusively on Components may only apply to Development and Production teams because the level of technical knowledge required to decompose and analyze parts of an AI system surpasses that of a typical Governance team member. For example, data scientists may be best prepared to perform documentation of data dependencies and pre-processing steps for *Model Info Sheets*, a framework that belongs to the Components category.[61] Alternatively, frameworks that fall solely into the Lifecycle category may be most relevant for Governance teams because buy-in from management may need to be secured before building, modifying, or removing an AI system. The planning, development, and deployment stages outlined in the Defense Innovation Unit's *Responsible AI Guidelines in Practice*, a Lifecycle framework, may require a Governance team stakeholder to sign off on the appropriateness of system success measures, approve rollback processes, assign accountability for change management, and make the final determination of whether the system is operationally useful.[62]

Figure 1: Governance Frameworks Make Up Most of the Matrix, and Their Distribution Across Utility Categories Differs from Development and Production Frameworks



	Development & Production	Governance
Components, Lifecycle, & Characteristics	0	4
Lifecycle & Characteristics	0	2
Components & Characteristics	2	4
Components & Lifecycle	3	5
Lifecycle	0	7
Components	4	0
Characteristics	5	9

Figure 1 displays the number of frameworks that fall into combinations of categories from the User and Utility dimensions. Governance and Development and Production frameworks are most heavily concentrated in the Characteristics category.

The following use cases illustrate how different users can leverage the matrix.

Use Case 1

A business executive works at an organization that develops AI-driven marketing solutions. He needs a framework that will help his organization adopt sustainable practices for building ethical AI. The executive can focus on those frameworks suited for Governance teams because coordinating activities across an organization is a governance responsibility. He is looking to establish end-to-end ethical design methods that are championed at all levels of the organization, from new employees to executives. These steps are not necessarily tied to a lifecycle stage or component of an AI system but are needed to build AI systems that achieve a level of legitimacy and buy-in within the organization. He can therefore constrain his search by focusing on frameworks that belong to the Characteristics category. A framework that he may pick is Salesforce's *AI Ethics Maturity Model*.[63]

Use Case 2

A data scientist is looking for a framework to help her responsibly document the datasets used by a machine learning model. The data scientist can focus on the frameworks best suited for Development and Production teams since data documentation is typically completed by those closest to the development and production of the machine learning model. She is interested in frameworks that specifically address data and emphasize the design phase of the AI system lifecycle, so she may narrow her search to frameworks in the Components and Lifecycle categories. A framework that she may pick is *Datasheets for Datasets*.[64]

Use Case 3

A lawyer within an organization developing AI-enabled cloud services is devising data protection policies for the organization. She is neither a part of a Development and Production team nor a Governance team, but her responsibility over compliance issues puts her in close coordination with both. She needs to liaise with the engineering team to ensure that technical practices for handling personal data are synchronized with regulations that the organization abides by. She may use a Governance framework, *The Aletheia Framework 2.0*, to structure initial questions for the engineering team, such as whether system architectures implement privacy by design or have the ability to update, amend, or remove an individual’s personal data if needed.[65] Since she is seeking information on the development and production of AI-enabled services to inform data protection policies, she can also reference Development and Production frameworks to more deeply explore the technical considerations at play.

Conclusion

Frameworks, no matter how thorough they may be, cannot take the place of deliberate consultation between interdisciplinary teams and impacted communities. And frameworks cannot definitively answer which problems an AI system should solve, what values the system should uphold, or how it will learn from and shape the environment in which it is served. However, process frameworks can raise important questions and provide flexible guidance around answering them. Unfortunately, these frameworks are currently scattered and difficult to compare. Organizations need a way of organizing and evaluating existing frameworks so they can effectively select ones that suit their needs.

This paper examined and organized more than 40 existing responsible AI frameworks to develop a user-centric approach to help organizations better understand, select, and implement responsible AI in a way that best fits their needs. Focusing on the user—Development and Production teams and Governance teams—supports accountability by tying procedures to organizational roles. The Utility dimension of the matrix developed in this paper enables a more precise selection of frameworks that meet the specific needs of framework users. Ultimately, the matrix helps organizations take the first step towards implementing responsible AI by putting the right resources in the hands of the people best equipped to do good with those resources.

Authors

Mina Narayanan is a research analyst at CSET working on AI Assessment.

Christian Schoeberl is a data research analyst at CSET working on AI Assessment.

Acknowledgments

The authors thank Heather Frase, Igor Mikolic-Torreira, Margarita Konaev, Karson Elmgren, Mia Hoffmann, Catherine Aiken, Emelia Probasco, Kathleen Curlee, Caroline Schuerger, Jessica Ji, Owen Daniels, and Luke Koslosky for their input on and review of the paper. The authors are grateful to Justin Lynch, Reva Schwartz, and Matthew Mahoney for their feedback and assistance, and to Lynne Weil for overseeing the paper's final revisions and publication.

Document Identifier: doi: 10.51593/20220029

Appendix A: Methodology

The matrix was populated with frameworks that were process-based, free, and openly available. However, the word “framework” is often generalized and connotes different meanings depending on the context in which it is used. For example, a software framework is understood as software with general functionalities that can be suppressed or extended. The processes within the frameworks that qualified for analysis had to enable systematic and actionable tasks, and a framework’s primary purpose had to be about implementing responsible AI within an organization.

The in- and out-citations of three prominent frameworks that met the aforementioned criteria were reviewed: *Datasheets for Datasets*, *Model Cards for Model Reporting*, and *FactSheets: Increasing Trust in AI Services through Supplier’s Declarations of Conformity*.[66] Since all three frameworks met the criteria, it was expected that in-citations (papers that referenced at least one of the three frameworks) and out-citations (papers referenced by at least one of the three frameworks) would point toward similar frameworks. References within the three frameworks were used to identify and review out-citations. Google Scholar was then consulted to identify and review the first 50 most-cited in-citations of the three frameworks.

In addition, AI newsletters published between January 2021 and June 2022 were searched for mentions of “framework,” and each search hit was checked against the criteria. *Import AI*, *ChinAI Newsletter*, *The AI Ethics Brief*, *The European AI Newsletter*, *The Machine Learning Engineer Newsletter*, *TechStream*, *The Batch*, and *policy.ai* were reviewed for frameworks.[67] Frameworks that colleagues recommended were also reviewed.

Exclusions to the review were nondescript lists of principles, ethical guidelines, and rights; guidance that involves AI but is not geared toward organizations building AI systems in-house, such as procurement guidance; checklists where processes are not organized into clear categories or themes; taxonomies; thought experiments; standalone datasets, metrics, or algorithms; frameworks for technology that is not AI; technical tools; mandatory guidance such as regulation, and standards published by standards-setting organizations such as the Institute of Electrical and Electronics Engineers and the International Organization for Standardization. Many standards live behind a paywall or are under development by a working group, and therefore are not free and openly available.

Appendix B: Examples of Frameworks that Belong to Components, Lifecycle, and Characteristics Categories

Components

Two examples of frameworks that are assigned to the Components category are *Model Cards for Model Reporting* and *Data Statements for Natural Language Processing*.[68] Model cards are short documents that describe key features of a trained machine learning model. They describe the provenance of the model, the data it relies on, and in-scope and out-of-scope usages. Model cards emphasize benchmarked evaluation across intersectional groups and encourage those responsible for the cards to articulate ethical considerations and quantify variability of metrics. Data statements focus on the annotators and subjects of datasets for natural language processing (NLP). Data statements therefore incorporate information about less visible communities that contribute to the finished data product. They aim to alleviate exclusion and bias in language technology by generalizing NLP research to other populations and obtaining consent from annotators and speakers for their data. Their schema includes a curation rationale, speaker demographics, annotator demographics, speech situation, and text characteristics.

Lifecycle

An example of a framework that belongs to the Lifecycle category is *Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems*.[69] It views machine learning as a four-step process that consists of commissioning, model building, decision-making, and investigation. Commissioning involves defining the problem that an algorithmic decision-making system solves, assessing its impact, and maintaining records related to procurement. The model building stage is divided into data collection, pre-processing, training, and testing in the form of verifiable claims. Manager oversight and knowledge differentials between data collectors and model developers are highlighted at this stage. Decision-making concerns the preparation of the system for deployment and the actual use of the system to make decisions and produce consequences. The final stage is investigation, where internal or external auditing is conducted to evaluate compliance and disclosures make information about the system available to others.

Characteristics

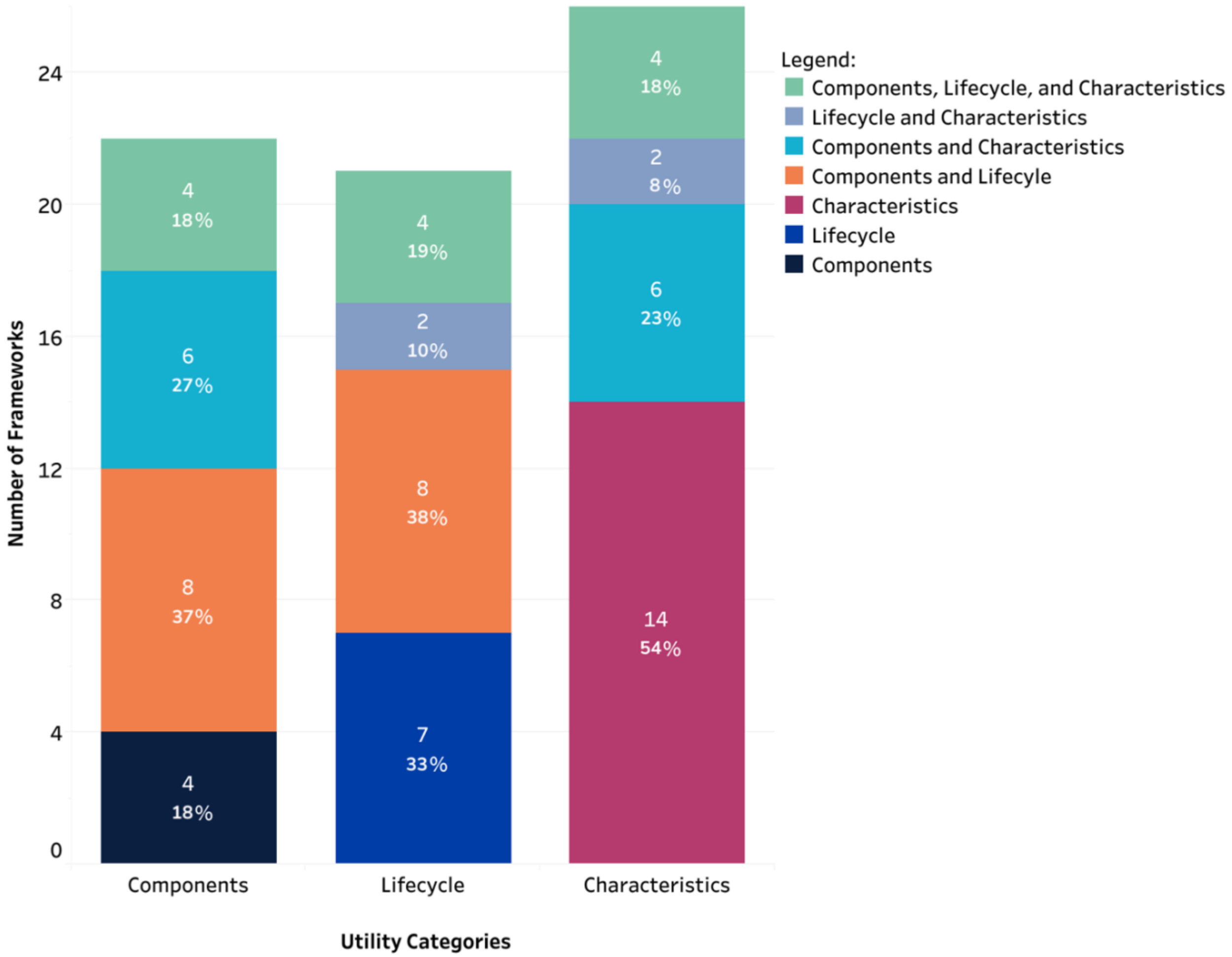
An example of a Characteristics framework is Google's *Responsible AI practices*.[70] Google's practices cover four primary principles: fairness, interpretability, privacy, and security. Fairness includes setting concrete goals, making sure representative datasets are used, checking for unfair biases by examining performance on subgroups, and stress-testing the AI system on difficult cases. Interpretability stresses the importance of user-friendliness and thoroughly testing the system. Privacy involves the responsible collection and processing of data and safeguarding models against privacy breaches. Finally, security is presented as identifying potential threats and developing approaches to combat those threats. The EU Commission's High-Level Expert Group on Artificial Intelligence's *Ethics Guidelines for Trustworthy AI* organizes guidelines under expansive categories that go beyond the more traditional principles of accountability and transparency and include environmental well-being and whether human agency and oversight were accounted for through methods such as impact assessments.[71]

Combination of Categories

Towards Accountability for Machine Learning Datasets is a framework that falls under the Components and Lifecycle categories because it conceptualizes data as an infrastructure and outlines five stages of specifications through worksheets.[72] The dataset requirements analysis stage is when use cases and stakeholder needs are elucidated. Dataset design answers how the requirements will be achieved and justifies design decisions, whereas dataset implementation prescribes documenting actions taken in the form of code comments, a dataset implementation diary, or an issue-tracking system. Dataset testing involves acceptance and adversarial testing, and the final stage of dataset maintenance accounts for contingencies such as dataset drift. *System Cards for AI-Based Automated Decision Systems* ties all three Utility categories together.[73] It presents a matrix with rows that map to the data, model, code, and AI system. The columns correspond to AI system lifecycle stages, which include development, assessment, and mitigation. Actions within each cell of the matrix refer to upholding principles such as fairness and privacy.

Figure B1: Some Utility Categories Are More Popular than Others Among Frameworks in the Matrix

Figure B1 shows how some Utility categories are more popular than others among frameworks in the matrix. The bar labeled Characteristics shows the number of frameworks that belong to the Characteristics category. Frameworks that belong to the Characteristics category may not belong to this category exclusively, as shown by the legend. The percentages that appear in the Characteristics bar represent shares of the total number of frameworks that fall under the Characteristics category, not shares of the total number of frameworks in the matrix. For example, 54% of Characteristics frameworks belong exclusively to the Characteristics category, whereas 23% of Characteristics frameworks belong to both the Components and Characteristics categories. The same applies for the Components and Lifecycle bars. Identically



colored segments of bars represent frameworks that are double- or triple-counted across categories in the figure.

Visualizing the distribution of frameworks this way is important because it may suggest focus areas where more frameworks are needed. By referencing Figure B1, organizations that develop their own framework can begin to gauge whether their framework contributes new information to guidance that already exists. Many frameworks that belong solely to the Characteristics category exist, likely because of their high signaling power and the ease with which organizations can espouse (but not necessarily implement) characteristics. If an organization decides to create a new framework, it might consider creating a Components or Lifecycle framework to maximize the marginal utility of the framework.

Appendix C: Framework Comparisons Using the Lifecycle and Characteristics Categories

Appendix C provides three examples for how a user could map common practices across frameworks that belong to the same Utility category to more precisely compare, select, and identify information gaps in frameworks. Table C1 depicts how the AI system lifecycle stages of three frameworks overlap.

Table C1: AI System Lifecycle Stages of Frameworks Overlap

Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems[74]	Commissioning	Model Building	Decision-Making	Investigation
Partnership on AI's ABOUT ML Reference Document[75]	Data & Model Specification	Data Curation, Model Training, Model Evaluation	Data & Model Integration	Data & Model Maintenance
BSA's Framework to Build Trust in AI[76]	Project Conception	Data Acquisition & Preparation, Model Definition & Testing	Preparing for Deployment & Use	

Several AI system lifecycle stages of *Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems*, Partnership on AI's *ABOUT ML Reference Document*, and *Confronting Bias: BSA's Framework to Build Trust in AI* share similar processes.

Note that the bars above each stage do not indicate the duration of a stage but instead compare the relative processes of different frameworks' stages. For example, the Investigation stage of *Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems* and the Data & Model Maintenance stage of

Partnership on AI's *ABOUT ML Reference Document* both share the processes of complying with regulation and continually assessing the suitability of technology.[77] While these two stages are not identical, they have similar processes and therefore share the same bar length and color scheme. On the other hand, the Data & Model Specification stages of Partnership on AI's *ABOUT ML Reference Document* include the processes of evaluating the potential impact of AI systems on communities from the Commissioning stage and choosing the model architecture and tests from the Model Building stage of *Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems*.[78] This mapping is represented by the bar for Data & Model Specification spanning both the length of the Commissioning bar and part of the Model Building bar.

Although the number and names of stages differ between the three frameworks, they span similar processes and therefore a user might simply pick the framework that most closely mirrors their existing project phases. This mapping supports the step of selecting a framework from a collection of frameworks that belong to the same category. It supports interoperability between Lifecycle frameworks by systematically identifying which stages denote similar processes but have different names.

Table C2 shows the lifecycle stages for two frameworks that focus on datasets. Note that most of the stages in *Datasheets for Datasets* and *Towards Accountability for Machine Learning Datasets*, except for Stages 4 and 5, overlap.[79] This means that for each of the overlapping stages, both frameworks articulate similar processes. For example, the second stages of both frameworks prescribe the documentation of design decisions, such as recording relationships between instances in a dataset or the use of sensitive data, and the justification of these decisions. As in Table C1, this does not imply that each stage in a framework, or corresponding stages in different frameworks, will take the same amount of time to complete or are identical. The tables illustrate how a user could leverage the matrix to identify frameworks that share similar high-level stages. A user could then feel empowered to combine frameworks or pick one upon closer inspection of the stages.

Table C2: Data Lifecycle Stages of Frameworks Overlap

	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6
Towards Accountability for Machine Learning Datasets[80]	Dataset Requirements	Dataset Design	Dataset Implementation	Dataset Testing		Dataset Maintenance
Datasheets for Datasets[81]	Motivation	Composition	Collection, Preprocessing, Cleaning, Labeling, Use		Distribution	Maintenance

Many data lifecycle stages of *Towards Accountability for Machine Learning Datasets* and *Datasheets for Datasets* share similar processes.

The gray boxes in Table C2 indicate a missing analogous stage in one of the frameworks. Since each framework is missing a stage relative to its counterpart, a user's calculus for selecting one framework over another may be more complicated than the previous example. Perhaps the user needs specific guidance on testing their datasets. In that case, *Towards Accountability for Machine Learning Datasets* is the appropriate choice.[82] Conversely, they may wish to build their own comprehensive data pipeline, in which case combining the two frameworks together would yield the most information.

Table C3 shows how the Characteristics category can also provide a standard way of comparing frameworks. Four frameworks that share responsible AI characteristics were selected to illustrate how a framework's inclusion or lack of processes for translating characteristics into practice can inform a user's actions. If a developer is concerned with addressing all four characteristics of accountability, transparency, fairness, and interpretability, she may familiarize herself with the practices in *AI Ethics Framework for the Intelligence Community* since it covers all characteristics of interest and is geared towards Development and Production teams.[83] However, if she wishes to compare a governance-oriented approach to AI transparency with industry-focused guidance around AI transparency, she may read *AI Ethics Impact Group: From Principles to Practice*, which is a framework suited for Governance teams that poses questions around transparency, and *Microsoft Responsible AI Standard, v2*, which describes steps that Development and Production teams can take to achieve transparency goals.[84] Like the two previous examples, Table C3 supports interoperability among frameworks by indicating which ones discuss certain characteristics and which do not.

Table C3: Frameworks Share Responsible AI Characteristics

	Accountability	Transparency	Fairness	Interpretability
Google's Responsible AI practices[85]			✓	✓
Microsoft Responsible AI Standard, v2[86]	✓	✓	✓	
AI Ethics Impact Group[87]	✓	✓		
AI Ethics Framework for the Intelligence Community[88]	✓	✓	✓	✓

Accountability, transparency, and fairness appear most frequently in the selected frameworks.

Endnotes

[1] The National Institute of Standards and Technology, *Trustworthy & Responsible AI Resource Center* (Washington, D.C.: Department of Commerce, 2023), https://airc.nist.gov/home.

[2] Organisation for Economic Co-operation and Development, *Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems* (Paris: OECD Publishing, 2021), https://doi.org/10.1787/008232ec-en.

[3] GitHub, Awesome AI Guidelines, accessed November 23, 2022, https://github.com/EthicalML/awesome-artificial-intelligence-guidelines.

[4] Confronting Bias: BSA's Framework to Build Trust in AI, BSA The Software Alliance, https://ai.bsa.org/confronting-bias-bsas-framework-to-build-trust-in-ai.

[5] The National Institute of Standards and Technology, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence* (Washington, D.C.: Department of Commerce, 2022), 3, https://doi.org/10.6028/NIST.SP.1270; Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, "Model Cards for Model Reporting," (paper presented at FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, Georgia, January 29–31, 2019), 220–229, https://dl.acm.org/doi/10.1145/3287560.3287596.

[6] "Microsoft Responsible AI Standard, v2," Microsoft, June 2022, https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf.

[7] Carol J. Smith, "Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development," arXiv preprint arXiv:1910.03515 (2019), https://arxiv.org/abs/1910.03515.

[8] Office of the Director of National Intelligence, *Artificial Intelligence Ethics Framework for the Intelligence Community* (Washington, D.C.: Office of the Director of National Intelligence, 2020), https://www.intelligence.gov/artificial-intelligence-ethics-framework-for-the-intelligence-community.

[9] "Responsible AI practices," Google, https://ai.google/responsibilities/responsible-ai-practices/.

[10] Colin Smith, Ewen Denney, and Ganeshmadhav J. Pai, "Hazard Contribution Modes of Machine Learning Components," (paper presented at AAAI Conference on Artificial Intelligence, New York, New York, February 7–12, 2020), https://ntrs.nasa.gov/citations/20200001851.

[11] Emily M. Bender and Batya Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," *Transactions of the Association for Computational*

Linguistics 6 (2018): 587–604, https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00041/43452/Data-Statements-for-Natural-Language-Processing.

[12] Shazia Afzal, C Rajmohan, Manish Kesarwani, Sameep Mehta, and Hima Patel, "Data Readiness Report," (paper presented at IEEE International Conference on Smart Data Services, Chicago, Illinois, September 5–10, 2021), https://ieeexplore.ieee.org/abstract/document/9592479.

[13] Mitchell et al., "Model Cards for Model Reporting."

[14] Sayash Kapoor and Arvind Narayanan, "Leakage and the Reproducibility Crisis in ML-based Science," arXiv preprint arXiv:2207.07048 (2022), https://arxiv.org/abs/2207.07048.

[15] Kathy Baxter, *AI Ethics Maturity Model*, Salesforce, https://www.salesforceairesearch.com/static/ethics/EthicalAIMaturityModel.pdf.

[16] Artificial Intelligence Impact Assessment, ECP Platform for the Information Society, 2022, https://ecp.nl/publicatie/artificial-intelligence-impact-assessment-english-version/.

[17] *The Aletheia Framework 2.0*, Rolls Royce, 2021, https://www.rolls-royce.com/~/media/Files/R/Rolls-Royce/documents/stand-alone-pages/aletheia-framework-booklet-2021.pdf.

[18] Information Commissioner's Office and The Alan Turing Institute, *Explaining decisions made with AI*, (Wilmslow, United Kingdom: Department for Digital, Culture, Media, and Sport, 2020), https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence/.

[19] World Economic Forum, *Empowering AI Leadership: An Oversight Toolkit for Boards of Directors* (Switzerland: World Economic Forum), https://express.adobe.com/page/RsXNkZANwMLEf/.

[20] High-Level Expert Group on Artificial Intelligence, *Ethics guidelines for trustworthy AI* (Brussels, Belgium: European Commission, 2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[21] Josef Baker-Brunnbauer, "TAII Framework for Trustworthy AI Systems," *ROBONOMICS: The Journal of the Automated Economy* 2 (December 2021): 17, https://journal.robonomics.science/index.php/rj/article/view/17.

[22] From Principles to Practice – An interdisciplinary framework to operationalise AI ethics, AIEI Group, https://www.ai-ethics-impact.org/en.

[23] "Ethics Framework," Machine Intelligence Garage, Digital Catapult, https://migarage.digicatapult.org.uk/ethics/ethics-framework/.

[24] Defense Innovation Unit, *Responsible AI Guidelines in Practice* (Washington, D.C.: Under Secretary of Defense for Research and Engineering, 2021), https://assets.ctfassets.net/3nanhbfkr0pc/acoo1Fj5uungnGNPJ3QWy/3a1dafd64f22efcf8f27380aafae9789/2021_RAI_Report-v3.pdf.

[25] "What Is the Cognitive Project Management for AI (CPMAI) Methodology?," Cognilytica, https://www.aidatatoday.com/what-is-the-cognitive-project-management-for-ai-cpmai-methodology/.

[26] Roel I.J. Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz, "Hard Choices in Artificial Intelligence: Addressing Normative Uncertainty through Sociotechnical Commitments," (paper presented at AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, New York, February 7–9, 2020), 242, https://dl.acm.org/doi/10.1145/3375627.3375861.

[27] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes, "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing," (paper presented at FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27, 2020), 33–44, https://dl.acm.org/doi/abs/10.1145/3351095.3372873.

[28] The National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (Washington, D.C.: Department of Commerce, 2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[29] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh, "Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems," (paper presented at FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Canada, March 3–10, 2021), 598–609, https://dl.acm.org/doi/10.1145/3442188.3445921.

[30] Confronting Bias: BSA's Framework to Build Trust in AI, BSA The Software Alliance.

[31] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford, "Datasheets for Datasets," *Communications of the ACM* 64, no. 12 (December 2021): 86–92, https://cacm.acm.org/magazines/2021/12/256932-datasheets-for-datasets/fulltext.

[32] Erin Lanus, Ivan Hernandez, Adam Dachowicz, Laura J. Freeman, Melanie Grande, Andrew Lang, Jitesh H. Panchal, Anthony Patrick, and Scott Welch, "Test and Evaluation Framework for Multi-Agent Systems of Autonomous Intelligent Agents," (paper presented at IEEE International Conference on System of Systems Engineering, Västerås, Sweden, June 14–18, 2021), https://ieeexplore.ieee.org/document/9497472.

[33] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell, "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure," (paper presented at FAccT '21: Proceedings of the 2021

ACM Conference on Fairness, Accountability, and Transparency, Canada, March 3–10, 2021), 560–575 https://dl.acm.org/doi/10.1145/3442188.3445918.

[34] Information Technology — Artificial Intelligence — Machine Learning (ML) model transparency, ATARC, 2020, https://atarc.org/project/information-technology-artificial-intelligence-machine-learning-ml-model-transparency/.

[35] "Responsible bots: 10 guidelines for developers of conversational AI," Microsoft, November 4, 2018, https://www.microsoft.com/en-us/research/uploads/prod/2018/11/Bot_Guidelines_Nov_2018.pdf.

[36] "Section 3: Preliminary Synthesized Documentation Suggestions," ABOUT ML Reference Document, Partnership on AI, September 7, 2021, https://partnershiponai.org/paper/about-ml-reference-document/7/#Section-3.

[37] Luciano Floridi, Matthias Holweg, Mariarosaria Taddeo, Javier Amaya Silva, Jakob Mökander, and Yuni Wen, "capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act," SSRN preprint SSRN:4064091 (2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091.

[38] AI Fairness Checklist, Microsoft, 2020, https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4t6dA.

[39] Michael Akinwumi, Lisa Rice, and Snigdha Sharma, "Purpose, Process, and Monitoring: A New Framework for Auditing Algorithmic Bias in Housing & Lending," National Fair Housing Alliance, February 17, 2022, https://nationalfairhousing.org/wp-content/uploads/2022/02/PPM_Framework_02_17_2022.pdf.

[40] Thomas Krendl Gilbert, Sarah Dean, Nathan Lambert, Tom Zick, and Aaron Snoswell, "Reward Reports for Reinforcement Learning," arXiv preprint arXiv:2204.10817 (2022), https://arxiv.org/abs/2204.10817.

[41] Personal Data Protection Commission, *Model Artificial Intelligence Governance Framework Second Edition* (Singapore: Personal Data Protection Commission, 2020), https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf.

[42] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney, "FactSheets: Increasing trust in AI services through supplier's declarations of conformity," *IBM Journal of Research and Development* 63 (September 2019): 6:1–6:13, https://ieeexplore.ieee.org/document/8843893.

[43] Information Commissioner's Office, *Guidance on the AI auditing framework* (Wilmslow, United Kingdom: Department for Digital, Culture, Media, and Sport, 2020), https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf.

[44] The National Institute of Standards and Technology, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*.

[45] Andrew Burt and Patrick Hall, "What to Do When AI Fails," O'Reilly, May 18, 2020, https://www.oreilly.com/radar/what-to-do-when-ai-fails/.

[46] Office of the Privacy Commissioner for Personal Data, *Guidance on the Ethical Development and Use of Artificial Intelligence* (Hong Kong: Office of the Privacy Commissioner for Personal Data, 2021), https://www.pcpd.org.hk/english/resources_centre/publications/files/guidance_ethical_e.pdf.

[47] U.S. Government Accountability Office, *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities* (Washington, D.C.: U.S. Government Accountability Office, 2021), https://www.gao.gov/products/gao-21-519sp.

[48] Furkan Gursoy and Ioannis A. Kakadiaris, "System Cards for AI-Based Decision-Making for Public Policy," arXiv preprint arXiv:2203.04754 (2022), https://arxiv.org/abs/2203.04754.

[49] Organisation for Economic Co-operation and Development, *OECD Framework for the Classification of AI Systems* (Paris: OECD Publishing, 2022), https://doi.org/10.1787/cb6d9eca-en.

[50] Artificial Intelligence & Technology Office, *DOE AI Risk Management Playbook (AIRMP)* (Washington, D.C.: Department of Energy 2022), https://www.energy.gov/ai/doe-ai-risk-management-playbook-airmp.

[51] Lanus et al., "Test and Evaluation Framework for Multi-Agent Systems of Autonomous Intelligent Agents."

[52] Smith et al., "Hazard Contribution Modes of Machine Learning Components."

[53] Carol J. Smith, "Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development."

[54] Organisation for Economic Co-operation and Development, *OECD Framework for the Classification of AI Systems*.

[55] Office of the Privacy Commissioner for Personal Data, *Guidance on the Ethical Development and Use of Artificial Intelligence*.

[56] Information Commissioner's Office, *Guidance on the AI Auditing Framework*.

[57] Afzal et al., "Data Readiness Report."

[58] "What Is The Cognitive Project Management For AI (CPMAI) Methodology?," Cognilytica.

[59] Artificial Intelligence & Technology Office, *DOE AI Risk Management Playbook (AIRMP)*.

[60] Information Commissioner's Office and The Alan Turing Institute, *Explaining decisions made with AI*.

[61] Kapoor et al., "Leakage and the Reproducibility Crisis in ML-based Science."

[62] Defense Innovation Unit, *Responsible AI Guidelines in Practice*.

[63] Baxter, *AI Ethics Maturity Model*.

[64] Gebru et al., "Datasheets for Datasets."

[65] *The Aletheia Framework 2.0*, Rolls Royce.

[66] Gebru et al., "Datasheets for Datasets"; Mitchell et al., "Model Cards for Model Reporting"; Arnold et al., "FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity."

[67] Jack Clark, *Import AI*, https://jack-clark.net/; Jeffrey Ding, *ChinAI Newsletter*, https://chinai.substack.com/; *The AI Ethics Brief*, Montreal AI Ethics Institute, https://brief.montrealethics.ai/; Charlotte Stix, *The European AI Newsletter*, https://www.europeanartificialintelligence.com/; *The Machine Learning Engineer Newsletter*, The Institute for Ethical AI & Machine Learning, https://ethical.institute/mle.html; *TechStream*, Brookings, https://www.brookings.edu/techstream/; *The Batch*, DeepLearning.AI, https://www.deeplearning.ai/the-batch/; *policy.ai*, Center for Security and Emerging Technology, https://cset.georgetown.edu/newsletters.

[68] Mitchell et al., "Model Cards for Model Reporting"; Bender et al., "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science."

[69] Cobbe et al., "Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems."

[70] "Responsible AI practices," Google.

[71] High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI.*

[72] Hutchinson et al., "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure."

[73] Gursoy et al., "System Cards for AI-Based Automated Decision Systems."

[74] Cobbe et al., “Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems.”

[75] “Section 3: Preliminary Synthesized Documentation Suggestions,” ABOUT ML Reference Document.

[76] Confronting Bias: BSA’s Framework to Build Trust in AI, BSA The Software Alliance.

[77] Cobbe et al., “Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems”; “Section 3: Preliminary Synthesized Documentation Suggestions,” ABOUT ML Reference Document.

[78] “Section 3: Preliminary Synthesized Documentation Suggestions,” ABOUT ML Reference Document; Cobbe et al., “Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems.”

[79] Gebru et al., “Datasheets for Datasets”; Hutchinson et al., “Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure.”

[80] Hutchinson et al., “Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure.”

[81] Gebru et al., “Datasheets for Datasets.”

[82] Hutchinson et al., “Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure.”

[83] Office of the Director of National Intelligence, *Artificial Intelligence Ethics Framework for the Intelligence Community*.

[84] AI Ethics Impact Group: From Principles to Practice, AIEI Group; “Microsoft Responsible AI Standard, v2,” Microsoft.

[85] “Responsible AI practices,” Google.

[86] “Microsoft Responsible AI Standard, v2,” Microsoft.

[87] AI Ethics Impact Group: From Principles to Practice, AIEI Group.

[88] Office of the Director of National Intelligence, *Artificial Intelligence Ethics Framework for the Intelligence Community*.