Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering

QINGHUA LU, LIMING ZHU, XIWEI XU, JON WHITTLE, DIDAR ZOWGHI, and AURELIE JACQUET, Data61, CSIRO, Australia

Responsible Artificial Intelligence (RAI) is widely considered as one of the greatest scientific challenges of our time and is key to increase the adoption of Artificial Intelligence (AI). Recently, a number of AI ethics principles frameworks have been published. However, without further guidance on best practices, practitioners are left with nothing much beyond truisms. In addition, significant efforts have been placed at algorithm level rather than system level, mainly focusing on a subset of mathematics-amenable ethical principles, such as fairness. Nevertheless, ethical issues can arise at any step of the development lifecycle, cutting across many AI and non-AI components of systems beyond AI algorithms and models. To operationalize RAI from a system perspective, in this article, we present an RAI Pattern Catalogue based on the results of a multivocal literature review. Rather than staying at the principle or algorithm level, we focus on patterns that AI system stakeholders can undertake in practice to ensure that the developed AI systems are responsible throughout the entire governance and engineering lifecycle. The RAI Pattern Catalogue classifies the patterns into three groups: multi-level governance patterns, trustworthy process patterns, and RAI-by-design product patterns. These patterns provide systematic and actionable guidance for stakeholders to implement RAI.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Machine learning**; • **Software and its engineering**;

Additional Key Words and Phrases: Responsible AI, ethical AI, trustworthy AI, AI governance, AI engineering, MLOps, software engineering, software architecture, pattern, best practice

ACM Reference format:
Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, Didar Zowghi, and Aurelie Jacquet. 2024. Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering. *ACM Comput. Surv.* 56, 7, Article 173 (April 2024), 35 pages.
https://doi.org/10.1145/3626234

1 INTRODUCTION

Artificial Intelligence (AI) has been transforming our society and listed as the top strategic technology in many organizations. Although AI has huge potential to solve real-world challenges, there are serious concerns about its ability to behave ethically and make decisions in a responsible way. Compared to traditional software systems, AI systems involve a higher degree of uncertainty and more ethical risk due to their dynamic, autonomous, and opaque decision making and historical data-dependent behaviors. **Responsible Artificial Intelligence (RAI)** refers to the ethical

Authors' address: Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Zowghi, and A. Jacquet, Data61, CSIRO, Australia; e-mails: {qinghua.lu, liming.zhu, xiwei.xu, jon.whittle, didar.zowghi, aurelie.jacquet}@data61.csiro.au.

0360-0300/2024/04-ART173
https://doi.org/10.1145/3626234

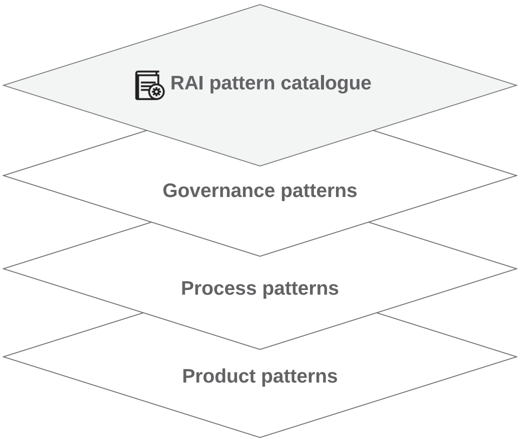


Fig. 1. Overview of the RAI pattern catalogue.

development of AI systems to benefit the humans, society, and environment. The concept of RAI has attracted significant attention from governments, organizations, companies, and societies. According to the 2022 Gartner CIO and Technology Executive Survey, 48% of organizations have already adopted or plan to adopt AI technologies within the next 12 months, whereas 21% of organizations have already deployed or plan to deploy RAI technologies within the next 12 months.[1] RAI has been widely considered as one of the greatest scientific challenges of our time and the key to unlock the market and increase the adoption of AI.

To address the RAI challenges, a number of AI ethics principles frameworks have been published recently [57], which AI systems are supposed to conform to. There has been a consensus made around the AI ethics principles [37]. A principle-based approach allows technology-neutral, future-proof, and context-specific interpretations and operationalization. However, without further best practice guidance, practitioners are left with nothing much beyond truisms. For example, it is a very challenging and complex task to operationalize the the human-centered value principle regarding how it can be designed, implemented, and monitored throughout the entire lifecycle of AI systems. In addition, significant efforts have been put on algorithm-level solutions which mainly focus on a subset of mathematics-amenable ethical principles (e.g., privacy and fairness). However, issues (including ethical issues) can occur at any step of the development lifecycle, crosscutting many AI, non-AI, and data components of systems beyond AI algorithms and models. To try to fill the principle-algorithmic gap, further guidance such as guidebooks,[2,3] questions to generate discussions [66, 67], checklists [44, 62], and documentation templates [1, 4, 52, 55, 94, 127] have started to appear. Those efforts tend to be ad-hoc sets of more detailed prompts for practitioners to think about all the issues and come up with their own solutions.

In this article, we therefore adopt a pattern-oriented approach and present an RAI Pattern Catalogue for operationalizing RAI from a system perspective. In software engineering, a pattern is a reusable solution to a problem that occurs commonly within a given context in software development [13]. Rather than staying at the ethical principle level or algorithm level, we focus on patterns that practitioners can utilize in practice to ensure that the developed AI systems are responsible throughout the entire software development lifecycle. As shown in Figure 1, the RAI

[1] https://www.gartner.com/en/articles/it-budgets-are-growing-here-s-where-the-money-s-going
[2] https://www.microsoft.com/en-us/haxtoolkit/
[3] https://pair.withgoogle.com/guidebook

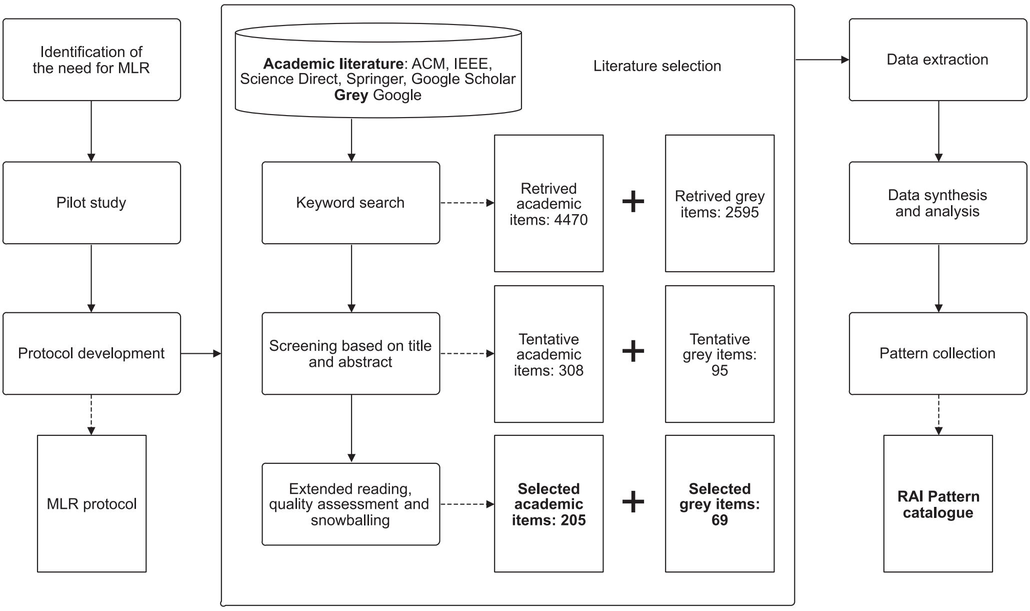

Identification of the need for MLR
Pilot study
Protocol development
MLR protocol
Academic literature: ACM, IEEE, Science Direct, Springer, Google Scholar
Grey Google
Literature selection
Keyword search
Retrived academic items: 4470
+
Retrived grey items: 2595
Screening based on title and abstract
Tentative academic items: 308
+
Tentative grey items: 95
Extended reading, quality assessment and snowballing
Selected academic items: 205
+
Selected grey items: 69
Data extraction
Data synthesis and analysis
Pattern collection
RAI Pattern catalogue

Fig. 2. Methodology.

Pattern Catalogue classifies patterns into three groups: (1) governance patterns for establishing multi-level governance for RAI, (2) process patterns for setting up trustworthy development processes, and (3) product patterns for building RAI-by-design paradigm into AI systems. These patterns are identified through conducting a systematic **Multivocal Literature Review (MLR)**. The full version of our RAI Pattern Catalogue can be accessed online.[4]

The remainder of the article is organized as follows. Section 2 introduces the methodology for building up the pattern catalogue. Section 3 presents the AI system stakeholders and governance patterns. Section 4 discusses the process patterns for each stage of the development lifecycle. Section 5 introduces the project patterns. Section 6 discusses the related work. Section 5 concludes the article.

2 METHODOLOGY

To build up an RAI Pattern Catalogue, we performed a systematic MLR to collect patterns. Figure 2 presents the research design and methodology. The high-level research question that has guided this research is this: "What RAI solutions can be identified?" The research question focuses on identifying the reusable patterns for RAI.

2.1 Data Preparation

The benefit of an MLR is to cover both academic literature and grey literature in the study. Grey literature is written by practitioners (e.g., governments, organizations, companies) and not published in books or scientific journals/conferences. However, grey literature can provide valuable insights on the state of practice and may include many industry solutions that are not discussed in

[4] Responsible AI Pattern Catalogue (https://research.csiro.au/ss/science/projects/responsible-ai-pattern-catalogue/). Each of the patterns is described following the traditional pattern structure (i.e., context, problem, solution, benefits, drawbacks, related patterns, known uses).

Table 1. Key and Supplementary Search Terms

Key Term	Supplementary TermsSupplementary Terms
AI	Artificial Intelligence, Machine Learning, ML
Responsible	Ethics, Ethical, Responsibility, Trust, Trusted, Trustworthiness, Trustworthy, Human Values, Wellbeing, Accountability, Accountable, Transparency, Transparent, Explainability, Explainable, Interpretability, Interpretable, Contestability, Contestable, Fairness, Fair, Reliability, Reliable, Safety, Safe, Privacy, Private, Security, Secure
Solution	Tactic, Practice, Process, Design, Architecture, Solution, Approach, Method, Mechanism, Tool, Toolkit

academic papers. Given the nature of patterns, we decided to also review grey literature to understand the state of the practice in the field of RAI and collect patterns from industry. In our MLR, we identify (1) relevant academic peer-reviewed academic literature and (2) relevant grey literature for this study.

2.2 Search Strategy

The study has been carried out separately for academic literature and grey literature. We adopted the **Systematic Literature Review (SLR)** guideline in the work of Kitchenham and Charters [60] to review the academic literature and used the guideline of Garousi et al. [39] to perform the grey literature review. The complete MLR protocol is available as online material.[5] Overall, we first tested different search strings in the five well-known search engines and evaluated the total number of studies retrieved, as well as their relevance. The evaluation involved cross checking the inclusion of known relevant literature. Once we determined the most effective search string, we proceeded to perform searches on Google Scholar. From the results, we randomly selected 20 papers and extracted the relevant answers for each of the research questions.

We determined the search strings by deriving relevant keywords from the research question. Before conducting the systematic search, we did a pilot study by experimenting with the search terms to compare the results. We used "AI," "Responsible," and "Solution" as the key terms and included synonyms and abbreviations as supplementary terms to increase the search results. We designed the search strings for each primary source to check the title. After completing the first draft of search strings, we examined the results of each search string against each database to check the effectiveness of the search strings. The finalized search terms are shown in Table 1. We use Australia's AI ethics principles [27] to identify the supplementary terms for "Responsible" as a close-enough representation of the many similar ones [37, 57] around the world. The eight AI ethics principles include *human, societal, and environmental wellbeing*; *human-centered values*; *fairness*; *privacy protection and security*; *reliability and safety*; *transparency and explainability*; *contestability*; and *accountability*. We mapped each individual term in the eight principles to its corresponding noun term and adjective forms. Furthermore, to encompass the relevant terms related to RAI, we expanded the mapping to include "responsible" as well as its variations, such as ethics, ethical, responsibility, trust, trusted, trustworthiness, and trustworthy. By doing so, we ensure comprehensive coverage of the terms relevant to RAI. The search strings and the respective paper quantities of the initial search for each primary source are listed in our MLR protocol. We applied the search string to both scholar search engines for academic literature and Google Search Engines for grey literature. The scholar search engines includeACM Digital Library, IEEE Xplore, Science Direct, Springer Link, and Google Scholar. The search period is up to July 31, 2022.

[5] https://drive.google.com/file/d/18Jiap714N1uprFVYU0jGmKSxKa-y2awJ/view?usp=sharing

We screened the initial results against inclusion and exclusion criteria. The inclusion criteria include the following: (1) a paper/article that presents a governance or process or design solution for RAI, (2) a paper/article that presents a tool or toolkit for developing RAI systems, and (3) a paper/article that is in the form of a published scientific paper or industry article. The exclusion criteria are as follows: (1) a paper/article that only discusses high-level principles or frameworks, (2) a paper/article that only focuses on algorithm-level techniques, (3) a paper/article that is not written in English, (4) a conference version of a study that has an extended journal version, and (5) Ph.D./Master's dissertations, tutorials, editorials and books.

The snowballing technique has been recommended and used in place of database searches in SLRs. For the academic literature, we identified a set of papers that serve as the starting point (i.e., seed set) for snowballing. The seed set papers were selected based on the source databases and Australia's AI ethics principles to cover various communities. For each of the five source databases, we selected one top-cited paper for RAI in general and each of the eight principles, respectively. For some principles, there is no paper found in one particular database. We only collected the grey literature from the first 10 Google pages. For the grey literature, snowballing is conducted if related RAI solutions are mentioned on the webpage. We finally identified 205 academic items and 69 grey items for the MLR. For the grey literature, we organized the RAI solutions according to the corresponding companies. For example, we found 13 RAI tools/solutions on Microsoft's website but only counted Microsoft as one grey item in our data extraction sheet and recorded a few patterns extracted from Microsoft's tools/solutions.

2.3 Data Extraction, Synthesis, and Analysis

To realize RAI from a software engineering perspective, we need to make both AI products and their development processes trustworthy and responsible. Additionally, compliance with AI standards and laws from a governance perspective is necessary. Thus, we classify the patterns into three categories: governance, process, and product. Not only should you use product patterns to enforce RAI principles directly in the product and verify/validate the product, but you should also use process and governance patterns to complement it further.

We extracted data and summarized findings from the selected academic and grey items based on the pre-defined research question. Based on the answers extracted for the research question, we identified different types of patterns. For example, there are a few papers using federated learning to deal with data privacy issues, thus "federated learner" is identified as a product pattern that can be built into the architecture of AI systems for continuous learning. Some of the solutions can be mapped to multiple levels. For example, the software bill of materials can be identified as an organization-level governance pattern that is interconnected with and supported by the product pattern "bill of materials registry." After identifying a pattern, we documented its details on our RAI Pattern Catalogue website according to the traditional pattern structure [74]: context, problem, solution, benefits, drawbacks, related patterns, and known uses. The known uses were found through data extraction and additional manual search. The pattern users and impacted stakeholders are identified based on (1) the AI software supply chain and ecosystem, (2) AI standards, and (3) our expertise and knowledge.

We also extract some general information, such as author name, organization, publication venue, and publication year. We performed a pilot study on 20 items to test the research question and the way to extract the required data. We stored all the extracted data in a spreadsheet for analysis.

3 GOVERNANCE PATTERNS

The governance for RAI systems can be defined as the structures and processes that are employed to ensure that the development and use of AI systems meet AI ethics principles. According to the

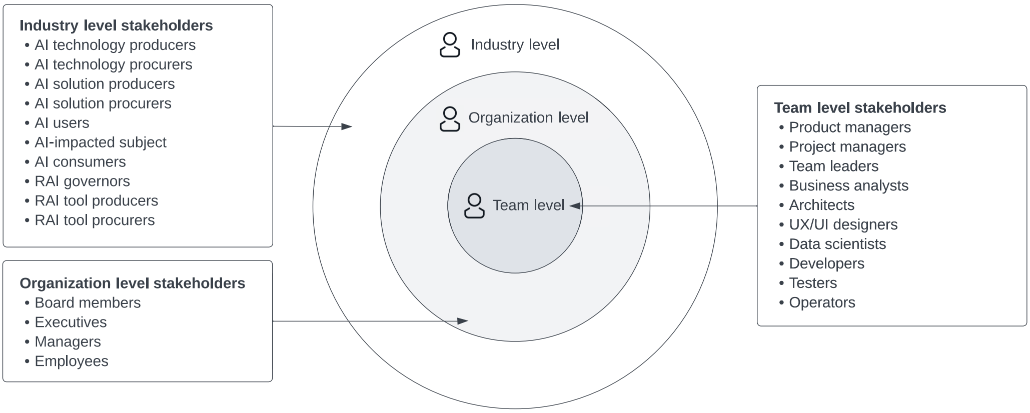


Fig. 3. Stakeholders for RAI governance.

structure of Shneiderman [104], governance can be built at three levels: industry level, organization level, and team level. As illustrated in Figure 3, we identified the stakeholders for RAI governance and classified them into three groups:

- *Industry-level stakeholders*:
  - *AI technology producers* develop AI technologies for others to build on top to produce AI solutions (e.g., parts of Google, Microsoft, IBM). AI technology producers may embed RAI in their technologies and/or provide additional RAI tools.
  - *AI technology procurers* procure AI technologies to build their in-house AI solutions (e.g., companies or government agencies buying/using AI platform/tools). AI technology procurers may care about RAI issues and embed RAI into their AI technology procurement process.
  - *AI solution producers* develop in-house/blended unique solutions on top of technology solutions and need to make sure the solutions adhere to RAI principles/standards/regulations (e.g., parts of MS/Google providing Office/Gmail “solutions”). They may offer the solutions to AI consumers directly or sell to others. They may use RAI tools (provided by AI technology producers or RAI tool producers) and RAI processes during their solution development.
  - *AI solution procurers* procure complete AI solutions (with some further configuration and instantiation) to use internally or offer to external AI consumers (e.g., a government agency buying from a complete solution from vendors). They may care about RAI issues and embed RAI into their AI solution procurement process.
  - *AI users* use an AI solution to make decisions that may impact on a subject (e.g., a loan officer or a government employee). AI users may exercise additional RAI oversight as the human-in-the-loop.
  - *AI-impacted subjects* are impacted by some AI-human dyad decisions (e.g., a loan applicant or a taxpayer). AI impacted subjects may care about RAI issues and contest the decision on dyad AI grounds.
  - *AI consumers* consume AI solutions (e.g., voice assistants, search engines, recommender engines) for their personal use (not affecting third parties). AI consumers may care about RAI issues and the dyad AI aspects of AI solutions.
  - *RAI governors* are those who set and enable RAI policies and controls within their culture. RAI governors could be functions within an organization in the preceding list or external (regulators, consumer advocacy groups, community).

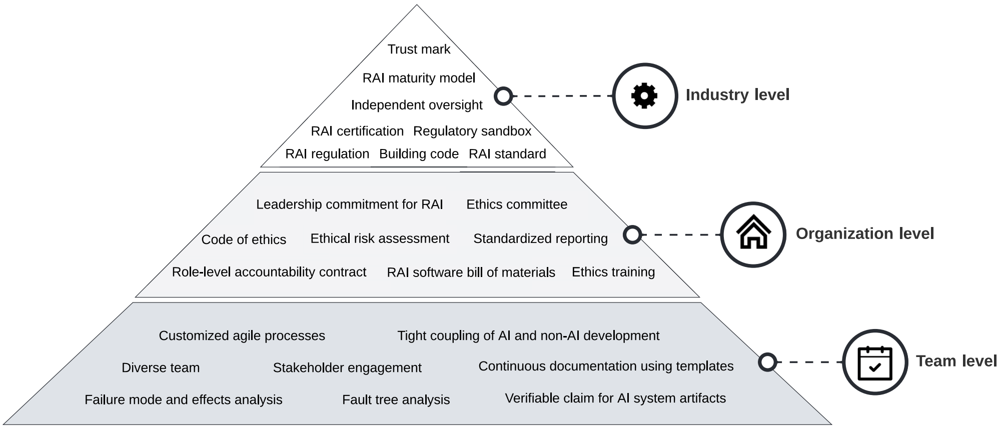


Fig. 4. Governance patterns for RAI.

  - *RAI tool producers* are technology vendors and dedicated companies offering RAI features integrated into AI platforms or AIOps/MLOps tools.
  - *RAI tool procurers* include any of the preceding stakeholders who may purchase or use RAI tools to improve or check solutions/technology's RAI aspects.
- *Organization-level stakeholders*:
  - *Management teams* include individuals at the higher level of an organization who are responsible for establishing RAI governance structure in the organization and achieving RAI at the organization level. The management teams include board members, executives and (middle-level) managers for legal, compliance, privacy, security, risk, and sustainability.
  - *Employees* are individuals who are hired by an organization to perform work for the organization and expected to adhere to RAI principles in their work.
- *Team-level stakeholders*:
  - *Development teams* include those who are responsible for developing and deploying AI systems, including product managers, project managers, team leaders, business analysts, architects, UX/UI designers, data scientists, developers, testers, and operators. The development teams are expected to implement RAI in their development process and embed RAI into the product design of AI systems.

As shown in Figure 4, we identify a set of governance patterns and classify them into industry-level governance patterns, organization-level governance patterns, and team-level governance patterns based on the governance structure of Shneiderman [104]. The target users of industry-level governance patterns are RAI governors, whereas the impacted stakeholders include AI technology producers and procurers, AI solution producers and procurers, and RAI tool producers and procurers. For the organization-level patterns, the target users are the management teams and the impacted stakeholders are employees, AI users, AI consumers, and AI-impacted subjects. The target users of team-level patterns are the development team, whereas the impacted stakeholders are AI users, AI consumers, and AI-impacted subjects.

3.1 Industry-Level Governance Patterns

3.1.1 RAI Regulation. Laws already apply to AI systems; however, the processes/requirements to ensure compliance are not always certain, and also some regulations may need to be updated (e.g., administrative law). There is an urgent need for clear guidance to ensure that AI systems are developed and used responsibly in compliance with existing and upcoming laws (e.g., discrimination law). RAI regulations are developed by governments in their jurisdiction to enable the trustworthy development of AI systems by industry [26, 53, 54, 91, 98, 104, 105]. Organizations will be required to ensure that they comply with the requirements of the EU AI Act when the applications fall into the high-risk category.[6] In the United States, the Algorithmic Accountability Act of 2022[7] was introduced in the Senate and House of Representatives, and an AI Bill of Rights[8] is under development by the White House Office of Science and Technology Policy. The aim of RAI regulations is to prevent illegal or negligent, malicious use of AI systems. However, there are many regulations in developments in each jurisdictions, which may cause an interoperability challenge for organizations. In addition, it usually takes a long time to enact AI regulations due to the lengthy consultation and approval process.

3.1.2 Regulatory Sandbox. To enable the trial of the innovative AI products in the market, a regulatory sandbox can be designed to allow testing the innovative AI products in the real world under relaxed regulatory requirements but with appropriate safeguards in place on a time-limited and small-scale basis [98]. An AI Regulatory Sandbox[9] is introduced in the EU's AI Act proposal submitted in 2021. The UK Information Commissioner's Office advised a Regulatory Sandbox [10] for utilizing personal data. The Australian Government released the Enhanced Regulatory Sandbox[11] for innovative financial services. AI products can enter the market under more flexible regulatory requirements in a faster pace and be tested in the real-world market to ensure that they are designed ethically. However, it might incur extra cost to apply for a regulatory sandbox. In addition, the AI products might not work well with large-scale deployment in different contexts.

3.1.3 Building Code. AI systems may have various degrees of risk depending on the design and application domains. To ensure that AI systems are trustworthy and meet certain minimum standards, building code can be designed to provide mandatory regulatory rules for authority parties (e.g., an independent oversight and advisory committee) to assess the compliance of AI systems before they are allowed to launch [104]. For example, IEEE has released a set of building codes for developing smart cities,[12] Medical Device Software Security,[13] and Power System Software Security.[14] The building code sets out clear compulsory regulatory requirements for developing AI systems. AI systems cannot be sold in the market until an approval is issued by the assessment authority.

3.1.4 RAI Standard. An AI system may use data or components from multiple jurisdictions that may have conflicting regulatory requirements on their usage. To enable interoperability between jurisdictions, RAI standards are developed to describe repeatable processes to develop and use AI systems responsibly that are recognized internationally and can be either mandated by law or by contract [53, 104, 105]. The ISO/IEC JTC 1/SC42 AI Technical Committee is developing the ISO/IEC 42001 IT-AI-Management System Standard,[15] which provides a pathway for the certification of AI systems and WG3 trustworthiness that covers risk management and bias.[16] IEEE has released

[6] https://artificialintelligenceact.eu
[7] https://www.congress.gov/bill/117th-congress/house-bill/6580/text?r=2&s=1
[8] https://www.wired.com/story/opinion-bill-of-rights-artificial-intelligence/
[9] https://www.eipa.eu/publications/briefing/sandboxes-for-responsible-artificial-intelligence/
[10] https://ico.org.uk/for-organisations/regulatory-sandbox/the-guide-to-the-sandbox/
[11] https://asic.gov.au/for-business/innovation-hub/enhanced-regulatory-sandbox/
[12] https://cybersecurity.ieee.org/blog/2017/10/04/building-code-for-the-internet-of-things/
[13] https://ieeecs-media.computer.org/media/technical-activities/CYBSI/docs/BCMDSS.pdf
[14] https://ieeecs-media.computer.org/media/technical-activities/CYBSI/docs/BCPSSS.pdf
[15] https://www.iso.org/standard/77304.html
[16] https://www.iso.org/committee/6794475.html

the Guide for Architectural Framework and Application of Federated Learning,[17] Standard for Technical Framework and Requirements of Trusted Execution Environment Based Shared Machine Learning,[18] and IEEE p7000 IEEE Standards for Model Process for Addressing Ethical Concerns During System Design.[19] Those AI standards provide repeatable processes and guidance for the use and development of AI systems that are recognized internationally.

3.1.5 RAI Maturity Model. Organizations can face challenges that can hurt their business if they are not aware of their RAI maturity. The RAI maturity model can be used to assess an organization's RAI capabilities and the degree of readiness to take advantage of AI based on a set of dimensions [5, 38, 104, 126]. The RAI maturity model can guide organizations on how to increase their RAI capabilities. The assessment results depend on the model quality, such as assessment dimensions and rating methods. There have been a few AI maturity models developed in industry, such as Gartner's AI Maturity Model,[20] Microsoft's AI Maturity Model,[21] and IBM's AI Maturity Framework.[22]

3.1.6 RAI Certification. AI is a high-stake technology that requires evidence to prove AI products' compliance with AI standards or regulations to operate in society. RAI certification can be designed to recognize that an organization or a person has the ability to develop or use an AI system in a way that is compliant with standards or regulations [19, 22, 26, 46, 72, 73, 104, 105, 124]. The Malta AI-ITA certification[23] is the world's first AI certification scheme for RAI systems. The DO-178C Certification[24] has been used to approve commercial software-based aerospace systems. Queen's University offers an executive education program on Principles of AI Implementation.[25] The evidence of compliance can be provided through RAI certification to improve human trust in AI systems. However, like other types of certificates, RAI certificates may be forged, which makes the verification of authenticity of certificates challenging. The certification process is usually complex, costly, and time consuming.

3.1.7 Trust Mark. Consumers in the market usually do not have professional knowledge about AI. To improve public confidence on AI and dispel their ethical concerns, the trust mark, a seal of endorsement, is easy to understand by all consumers and can be used to inform consumers about the AI system. The trust mark is important for small companies which are often not well known in the AI market. However, consumers may not trust that the AI systems with a trust mark perform more responsibly than those without one. There have been several trust marks designed for responsible use of data, such as the Australian Data and Insights Association Trust Mark,[26] the New Zealand Privacy Commissioner's Privacy Trust Mark,[27] and Singapore's Data Protection Trustmark (DPTM).[28]

3.1.8 Independent Oversight. Decisions made by AI systems may lead to severe failures due to its autonomous decision-making process. To audit AI systems and investigate failures in a trusted

[17] https://standards.ieee.org/ieee/3652.1/7453/
[18] https://standards.ieee.org/ieee/2830/10231/
[19] https://ethicsinaction.ieee.org/p7000/
[20] https://www.gartner.com/smarterwithgartner/the-cios-guide-to-artificial-intelligence
[21] https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4DIvg
[22] https://www.ibm.com/downloads/cas/OB8M18WR
[23] https://www.mdia.gov.mt/innovative-technology-arrangement-guidelines/
[24] https://my.rtca.org/nc__store?search=do-178c
[25] https://smith.queensu.ca/ConversionDocs/Execdev/Trusted_Data_AI_Canadian_Business.pdf
[26] https://dataandinsights.com.au/trust-mark/
[27] https://www.privacy.org.nz/resources-2/applying-for-a-privacy-trust-mark/
[28] https://www.pdpc.gov.sg/overview-of-pdpa/data-protection/business-owner/data-protection-trustmark

way, independent oversight can be conducted by independent oversight boards that consist of experts who are knowledgeable to perform the review and have no conflict of interest with the reviewed organizations [76, 104, 105, 128]. The U.S. National Transportation Safety Board[29] investigates every civil aviation accident. The U.S. National Artificial Intelligence Advisory Committee[30] advises on AI-related issues. Independent oversight provides a trusted review infrastructure to gain public confidence. Planning oversight provides early feedback on the new development proposals. Failures of independent oversight could happen due to lack of sufficient independence.

3.2 Organization-Level Governance Patterns

3.2.1 Leadership Commitment for RAI. The management teams need to understand the values, cost, and risk for adopting AI in an organization. Commitment needs to be made by the management team to build an RAI culture within an organization [104]. Leadership commitment is achieved by the management team dedicating their time and efforts on establishing ethics principles and governance structure (e.g., appointment of a chief RAI officer, RAI advisory boards) [108], as well as incorporating RAI into an organization's values, vision, mission [105], board strategy planning, executives' performance reviews [98], audit and risk committee's scope [54], and ESG commitments. Leadership commitment enables organizational culture on RAI and visible sponsorship to build RAI capability. IBM has established an AI ethics board[31] to support a culture of RAI throughout IBM. Axon has assembled an independent AI ethics board[32] to provide guidance on AI system development. Schneider Electric has appointed its first chief AI officer[33] to advance its AI strategy.

3.2.2 Ethics Committee. Organizations need to build capability incorporating multiple areas of expertise to address RAI issues. An AI ethics committee is an AI governance body that is established to develop standard processes for decision making, as well as to approve and monitor AI projects [19, 73]. Adobe has created an AI ethics committee[34] that includes experts from different backgrounds. Sony has established an AI ethics committee[35] to ensure the ethically development of AI systems. The ethics committee provides feedback and guidance to the project team after reviewing the proposal. However, the committee might not have the expertise to review a particular case, which might cause bias issues.

3.2.3 Code of Ethics for RAI. AI may make wrong decisions or behave inappropriately (e.g., impact human lives or buy the wrong product). To guide AI-related activities in an organization, a code of ethics is a set of rules that employees should uphold when developing an AI system [48, 53, 73, 104]. AAAI has issued the Code of Professional Ethics and Conduct[36] for all members. Bosch sets out a code of ethics[37] to establish guidelines for the development of AI. BMW has released a code of ethics for AI.[38] A code of ethics provides employees with the same concrete rules on

[29] https://www.ntsb.gov/

[30] https://www.ai.gov/naiac/#ABOUT-NAIAC

[31] https://www.ibm.com/au-en/artificial-intelligence/ethics

[32] https://www.axon.com/company/ai-and-policing-technology-ethics

[33] https://www.iteuropa.com/news/schneider-electric-appoints-new-caio-and-opens-new-ai-hub

[34] https://www.adobe.com/about-adobe/aiethics.html

[35] https://www.sony.com/en/SonyInfo/sony_ai/responsible_ai.html

[36] https://www.aaai.org/Conferences/code-of-ethics-and-conduct.php

[37] https://www.bosch-ai.com/industrial-ai/code-of-ethics-for-ai/

[38] https://www.press.bmwgroup.com/global/article/detail/T0318411EN/seven-principles-for-ai-bmw-group-sets-out-code-of-ethics-for-the-use-of-artificial-intelligence

developing AI systems, but it relies on individuals to do the right thing with limited monitoring and enforcement.

3.2.4 Ethical Risk Assessment. Although there are increasing concerns on AI ethics, RAI regulation is still at a very early stage. To assess the ethical risks associated with AI systems, an organization needs to extend the existing IT risk framework or design a new one to cover AI ethics [4, 14, 25, 31, 54, 64, 73, 94, 104]. The ISO/IEC JTC 1/SC 42 committee is developing ISO/IEC 23894 on Artificial Intelligence and Risk Management.[39] NIST released the initial draft of the AI Risk Management Framework that provides a standard process for managing risks of AI systems.[40] The Canadian government has released the Algorithmic Impact Assessment tool to identify the risks associated with automated decision-making systems.[41] The Australian NSW government is mandating all of its agencies that are developing AI systems to go through the NSW AI Assurance Framework.[42] Singapore launched the AI Verify Toolkit to test RAI.[43] UK ICO released the AI and Data Protection Risk Toolkit,[44] which is built up on their guidance for organizations using AI systems. Although ethical risk assessment has the potential to prevent the majority of incidents and increase awareness of RAI, it is often a one-off type of risk assessment with subjective judgment on measurement [95].

3.2.5 Standardized Reporting. Standardized reporting is essential to address the opaque black box issue of AI systems. Organizations should set up standardized processes and templates for informing the development process and product design of AI systems to different stakeholders (e.g., AI governors, users, consumers) [100]. RAI regulations may request such obligations to ensure the transparency and explainability of AI systems. The Cyberspace Administration of China published transparent disclosure requirements for online service providers.[45] The service providers are requested to file with the regulators (i.e., AI governors) for impact assessment when realizing new services. In addition, the online services must inform users when AI is being used to recommend content to them and explain the purposes and design of recommended systems. In the EU's AI Act,[46] the incidents of AI systems are required to be reported and disclosed by AI system providers (i.e., AI technology or solution producers). Helsinki[47] and Amsterdam[48] released AI registers describing where and how the two cities are using AI, how AI is built, which data and algorithms are used, how the applications impact the citizens' daily lives, and the development team's contact information.

3.2.6 Role-Level Accountability Contract. It is necessary that organizations have an appropriate approach to enable accountability throughout the entire lifecycle of AI systems. Role-level accountability can be established through formal contracts to define the boundary of responsibility and identify who should be held accountable when an AI system misbehaves [128]. For example, Australia's National Data Commissioner created a data sharing agreement template for using

[39] https://www.iso.org/standard/77304.html
[40] https://www.nist.gov/itl/ai-risk-management-framework
[41] https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html
[42] https://www.digital.nsw.gov.au/policy/artificial-intelligence/nsw-ai-assurance-framework
[43] https://file.go.gov.sg/aiverify.pdf
[44] https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/
[45] http://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm
[46] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
[47] https://ai.hel.fi/en/ai-register/
[48] https://algoritmeregister.amsterdam.nl/en/ai-register/

Australian Government data.[49] Developers primarily focus on the technique aspects of AI systems and may not be familiar with the ethical principles. Role-level accountability contracts make the developers keep ethics in mind at every step, but they may create stress for employees at all levels within an organization.

3.2.7 RAI Software Bill of Materials. From a software supply chain angle, the development of AI systems involves a complex and dynamic software supply chain. Many organizations procure AI technologies/solutions to build their AI systems. The AI systems are often assembled by using commercial or open source AI and/or non-AI components from third parties. Despite cost efficiency, the underlying security and integrity issues of the third-party components have attracted significant attention. According to Sonatype's report, 2021 State of the Software Supply Chain,[50] software supply chain attacks increased 650% in 2021, whereas it was 430% in 2020. The RAI software bill of materials keeps a list of components used to create an AI software product, which can be used by AI solution procurers and consumers to check the supply chain details of each component of interest and make buying decisions [12]. The supply chain details should at least include component name, version, supplier, dependency relationship, author of software bill of materials data, and timestamp [86]. This provides traceability and transparency about components and allows AI solution procurers and consumers to easily check component information (e.g., supply chain details and context information) and track ethical issues. The RAI software bill of materials enables faster vulnerability identification but may need to be updated frequently since AI systems may evolve over time. Dependency-Track[51] is widely used by practitioners to track components' supply chain information and identify known vulnerabilities. Software Package Data Exchange (SPDX)[52] and CycloneDX[53] are two standards for exchanging software bill of material information for security analysis.

3.2.8 Ethics Training. It is urgent that the employees of organizations begin to think through the potential implications of AI on their work and make ethical choices during the development and use of AI systems. Ethics training provides employees with knowledge on how to deal with ethical issues during development [6, 31, 53, 59, 97, 98, 104, 105, 128]. MIT offers a 3-day course, "Ethics of AI: Safeguarding Humanity,"[54] introducing the ethics of AI development and deployment. The University of Technology Sydney (UTS) designed a short course, "Ethical AI: from Principles to Practice,"[55] for business executives. The University of Helsinki created a free online course, "The Ethics of AI,"[56] for anyone interested in AI ethics. Halmstad University provides a short course on critical design and practical ethics for AI.[57] Ethics training can improve organizational awareness on RAI and sharpen the employees' RAI skills. However, RAI covers a broad range of knowledge and skills, and ethics training may only offer a subset of knowledge and skills within limited time.

3.3 Team-Level Governance Patterns

3.3.1 Customized Agile Process. Agile development has been increasingly adopted by organizations to incrementally and iteratively develop software systems, including AI systems. However,

[49] https://www.datacommissioner.gov.au/data-management/data-sharing-agreement
[50] https://www.sonatype.com/resources/state-of-the-software-supply-chain-2021
[51] https://dependencytrack.org/
[52] https://spdx.dev/
[53] https://cyclonedx.org/
[54] https://professional.mit.edu/course-catalog/ethics-ai-safeguarding-humanity
[55] https://open.uts.edu.au/uts-open/study-area/Technology/ethical-ai-from-principles-to-practice/
[56] https://ethics-of-ai.mooc.fi/
[57] https://learning4professionals.se/showCourse/158/Critical_Design_and_Practical_Ethics_for_AI

the existing agile development methods mainly focus on business value and largely neglect the AI ethics principles. To address ethical issues in the AI system development process, agile methods need to be extended and customized to allow consideration of ethics principles. Extension points could be artifacts, roles, ceremonies, practices, and culture [51]. Microsoft's Azure DevOps allows the customization of inherited processes.[58] Atola Technology provides a customized agile methodology that contains different development practices.[59] Apptio Targetprocess is a web-based visual tool for managing projects with flexibility at various levels.[60]

3.3.2 Tight Coupling of AI and Non-AI Development. AI system development involves the development of both AI and non-AI components with rapid iterations. This requires more frequent integration of AI and non-AI components. Compared with non-AI components, the development of AI components that support the AI model pipeline is more experimental with still limited methodological support and mostly done by data scientists and data engineers who are not familiar with software engineering. To bridge the methodological gap between AI and non-AI development, both the AI team and the non-AI team need to be clear about what exactly is being delivered by a project and share the same sprints and use a common co-versioning registry to track the progress [71]. The close coupling of AI and non-AI development results in improved trust within the project team and better communication on both system-level and model-level ethical requirements. The challenge for the tight coupling might be that the non-AI component development is application centric, whereas the AI component development is mostly data centric. There have been a few attempts in industry on continuously integrating AI components/models into the software, such as Microsoft Team Data Science Process,[61] Amazon SageMaker Pipelines,[62] and Azure Pipelines.[63]

3.3.3 Diverse Team. AI pipelines are built by humans and thus may imply bias (e.g., racism and sexism) and produce discriminating results. In addition, the code of AI systems is written by developers who are primarily focused on technical aspects. Building a diverse project team can effectively eliminate bias and improve diversity and inclusion in AI systems [31, 72, 131]. The diversity can be across gender, race, age, sexual orientation, expertise, and so on. A diverse team can drive creative thinking for greater innovation, but communication could become challenging due to different background and preference.[64] Google published the 2022 Diversity Annual Report,[65] which introduces the actions to build an inclusive workplace. Microsoft aims to integrate diversity and inclusion principles into their organization.[66] Meta has been working on creating diverse and inclusive work communities.[67]

3.3.4 Stakeholder Engagement. Stakeholders may have various ethical concerns about the development and use of AI systems. Keeping stakeholder engagement throughout the AI project is essential to building AI systems responsibly. Stakeholder engagement allows AI systems to better reflect their stakeholders' needs and expectations [17, 104, 129, 131]. There are various manners to engage stakeholders, such as interviews, online and offline meetings, project planning/review,

[58] https://docs.microsoft.com/en-us/azure/devops/organizations/settings/work/inheritance-process-model?view=azure-devops&tabs=agile-process
[59] https://www.airtable.com/universe/exp4OppRObzXbhOQE/custom-agile-methodology-by-atola
[60] https://www.apptio.com/products/targetprocess/
[61] https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview
[62] https://aws.amazon.com/sagemaker/pipelines/?nc1=h_ls
[63] https://www.azuredevopslabs.com/labs/vstsextend/aml/
[64] https://futureofworking.com/11-advantages-and-disadvantages-of-diversity-in-the-workplace/
[65] https://about.google/belonging/at-work/
[66] https://careers.microsoft.com/us/en/diversityandinclusion
[67] https://www.workplace.com/diversity-and-inclusion

participatory design workshops, and crowd sourcing. Stakeholders may help the project team identify potential ethical risks before they become threats, but there maybe conflicting opinions from different stakeholders. The Association for Project Management published 10 stakeholder engagement principles.[68] The Australian Public Service Commission released stakeholder engagement guidelines.[69] Deloitte published a report on stakeholder engagement.[70]

3.3.5 *Continuous Documentation Using Templates.*

Developers primarily focus on the code and often neglect updating the documentation during rapid iterations. The project teams need to create and continuously update documentations for the key artifacts of AI systems that may lead to ethical issues, such as data and models. Continuous documentation using templates helps track the evolution of artifacts and clarify the context in which AI systems are trustworthy [1, 4, 52, 55, 94, 127]. Google's Model Cards[71] enables transparent model reporting on model provenance and ethical evaluation [81, 119]. Microsoft's datasheets for datasets[72] tool allows every dataset to be accompanied with a datasheet document [41]. IBM's AI service FactSheets[73] maintains AI services' performance, safety, security, and provenance information [9]. Meta's method cards provide prescriptive model specification templates that provide guidance on how to mitigate potential issues [1, 2].

3.3.6 *Failure Mode and Effects Analysis.*

Ethical defects in AI systems are often detected through extensive simulation and testing in the later stages of development. However, this may lead to significant delays to timelines and additional development cost. **Failure Mode and Effects Analysis (FMEA)** is a bottom-up risk assessment method that can be used to identify ethical risks and calculate their priorities at the beginning of the development process [30]. FMEA was originally proposed in U.S. Armed Forces Military Procedures document MIL-P-1629 in 1949.[74] Ford Motor Company first introduced FMEA to the automotive industry in the mid-1970s.[75] FMEA has been extended and adopted by Toyota's Design Review Based on Failure Modes (DRBFM)[76] for assessing potential risk and reliability for automotive and non-automotive applications. FMEA replies on experts to apply their professional knowledge and experience to the ethical risk assessment process. In addition, FMEA is better suited for bottom-up analysis and not able to detect system-level complex ethical failures.

3.3.7 *Fault Tree Analysis.*

Undesired system behaviors or decisions could lead to serious consequences and even cause loss of human lives. **Fault Tree Analysis (FTA)** [30] can be used to describe how system-level ethical failures are led by small ethical failure events through an analytical graph (i.e., fault tree). The development team can easily capture how ethical failures propagate in the AI system. FTA can be done during the design or operation stage to anticipate the potential ethical risks and to recommend mitigation actions. FTA was first introduced by Bell Laboratories in 1962 to assess the safety of a missile launch control system.[77] Boeing started using FTA to de-

[68] https://www.apm.org.uk/resources/find-a-resource/stakeholder-engagement/key-principles/
[69] https://www.apsc.gov.au/initiatives-and-programs/workforce-information/taskforce-toolkit/stakeholder-engagement
[70] https://www2.deloitte.com/content/dam/Deloitte/za/Documents/governance-risk-compliance/ZA_StakeholderEngagement_04042014.pdf
[71] https://modelcards.withgoogle.com/about
[72] https://www.microsoft.com/en-us/research/project/datasheets-for-datasets/
[73] https://www.ibm.com/blogs/research/2018/08/factsheets-ai/
[74] https://web.archive.org/web/20110722222459/https://assist.daps.dla.mil/quicksearch/basic_profile.cfm?ident_number=37027
[75] https://fsp.portal.covisint.com/documents/106025/14555722/FMEA+Handbook+v4.2/4c14da5c-0842-4e60-a88b-75c18e143cf7?version=1.0
[76] https://www.sae.org/standards/content/j2886_201303/
[77] https://www.osti.gov/servlets/purl/1315144

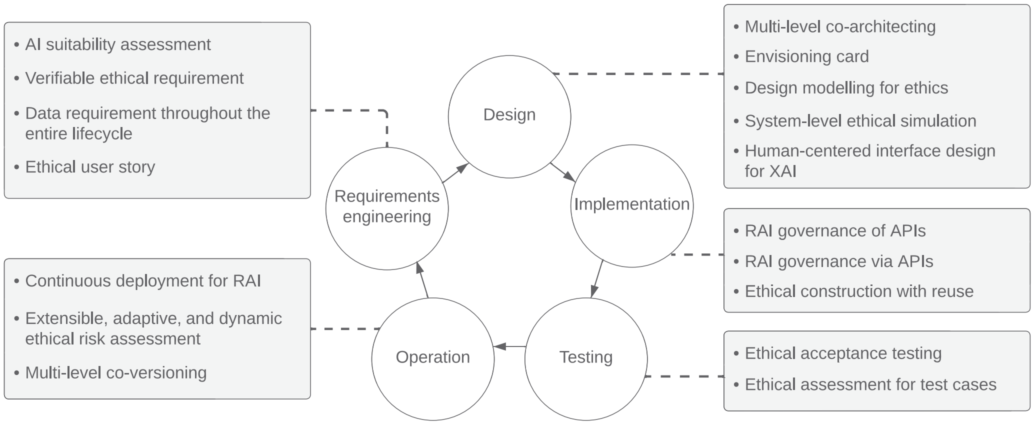

• AI suitability assessment
• Verifiable ethical requirement
• Data requirement throughout the entire lifecycle
• Ethical user story
Design
Requirements engineering
Implementation
Operation
Testing
• Multi-level co-architecting
• Envisioning card
• Design modelling for ethics
• System-level ethical simulation
• Human-centered interface design for XAI
• RAI governance of APIs
• RAI governance via APIs
• Ethical construction with reuse
• Continuous deployment for RAI
• Extensible, adaptive, and dynamic ethical risk assessment
• Multi-level co-versioning
• Ethical acceptance testing
• Ethical assessment for test cases

Fig. 5. Process patterns for RAI system development.

sign civil aircrafts in 1966.[78] FTA was included in the U.S. Army Materiel Command's Engineering Design Handbook on Design for Reliability.[79] FTA assists in analyzing the ethical issues related to AI system artifacts and prioritizes the issues to address that contribute to an ethical risk. However, it is complex to use for large system analysis, which may involve many ethical events and gates. In addition, time can hardly be captured in FTA.

3.3.8 *Verifiable Claim for AI System Artifacts.* The potential users of AI systems need methods for assessing an AI system's ethical properties and comparing the system to other systems. A verifiable claim platform can be built to support developers in making claims on ethical properties [40] and conducting the verification [124]. Such platform must consider the disparity of the stakeholder's views. For example, developers might focus on reliability, whereas users might be interested in fairness. A verifiable claim is a statement about an AI system or an artifact (e.g., model or dataset) that is substantiated by a verification mechanism. The platform itself provides management capabilities such as claim creation and verification, access control, and dispute management. The W3C Verifiable Claims Working Group aims to make expressing and exchanging claims.[80] The Open Web Application Security Project has published a Verifiable Claims documentation.[81] The Ethereum Verifiable Claims is a method for off-chain variable claims.[82]

4 PROCESS PATTERNS

In this section, we discuss the process patterns that can be incorporated into RAI system development processes. The process patterns are reusable methods and best practices that can be used by the development team during the development process. Figure 5 illustrates the process patterns collected for each stage of the development process.

4.1 Requirement Engineering

4.1.1 *AI Suitability Assessment.* AI has a huge potential to provide effective solutions to tackle critical problems. However, it does not necessarily add value to every software system. Before

[78] https://apps.dtic.mil/sti/citations/AD0847015
[79] https://apps.dtic.mil/sti/pdfs/ADA026006.pdf
[80] https://www.w3.org/2017/vc/WG/
[81] https://owasp.org/www-pdf-archive//OWASP-Austin-Mtg-2018Jan-CryptoParty-Dave-Sanford.pdf
[82] https://eips.ethereum.org/EIPS/eip-1812#ethereum-verifiable-claims

starting to build a software system with AI, the development team first needs to identify the right problem to solve and the corresponding user needs. Once the problem is found and the environment where the system will be situated fully explored, the development team needs to analyze whether the system and the users benefit from AI of if they are potentially degraded by AI [89]. It is essential to make sure that AI adds value to the design. Oftentimes, a heuristic-based design may be easier and cheaper to develop and may work better than an AI-based design in terms of predictability and transparency. AI suitability assessment can help the development team understand whether AI can add unique value to the design but may incur additional cost and require extra resources.

4.1.2 Verifiable Ethical Requirement. The development of AI systems needs to adhere to AI ethics principles which are generally abstract and domain agnostic. Ethical requirements need to be derived from the AI ethics principles to fit into a specific domain and system context [16, 49, 93, 118, 128]. Every ethical requirement specified in a requirements specification document should be put into a verifiable form (i.e., with acceptance criteria). This means that a person or machine can later check that the AI system meets the ethical requirements that are derived from AI ethics principles and grounded in users' needs. Vague or unverifiable statements should be avoided [110]. If there is no way to determine whether the AI system meets a particular ethical requirement, then this ethical requirement should be revised or removed. Ethical risk can be reduced via considering ethical requirements from the beginning of the development process and explicitly verifying ethical requirements. Some ethical principles/requirements may not be easily quantitatively validated [128], such as human-centered values. There may be tradeoffs between some ethical principles or requirements. The current practice to deal with the tradeoffs is usually the developers following one principle while overwriting the others rather than building balanced tradeoffs through patterns.

4.1.3 Data Requirements throughout the Entire Lifecycle. The quality of an AI model is largely dependent on the quality of the data used to train or evaluate. The lifecycle of data consists of several phases, including data collection, cleaning, preparation, validation, analysis, and termination. Unfortunately, the scope of data requirements [104, 118] often focuses on the data analysis phase and largely neglects the other key phases in the data lifecycle. This may lead to downstream ethical concerns such as AI model reliability, accountability, and fairness. AI systems can hardly be trusted when the data lifecycle is poorly managed. Data requirements need to be listed explicitly and specified throughout the data lifecycle (i.e., collection, cleaning, preparation, validation, analysis, and termination), taking into account ethical principles and involved stakeholders (i.e., data providers, data engineers, data scientists, data consumers, data auditors). Data requirements can be managed through data requirements specification. The specification could include detailed requirements for each phase in the data lifecycle, such as data collection requirements including data sources and collection methods. Google has created a template for dataset requirements specification [52].

4.1.4 Ethical User Story. Requirements elicitation methods are needed to collect detailed ethical requirements from stakeholders to capture AI ethics principles. In agile processes, ethical user stories [43, 93] can help the development team elicit ethical requirements for AI systems and implement AI ethics principles from the early stage of development. Ethical user stories are created to serve as items of the product backlog that is to be worked on by the development team in iterations (i.e., sprints). Card-based toolkits can be used to list questions related to AI ethics principles. The answers to those questions are integrated into ethical user stories to be included in sprint backlogs. The development team or users can write ethical user stories on cards or notes using a pre-defined template and assign them to different sprints based on priority. Ethical user stories make ethical requirements traceable both backward and forward, but they are difficult to scale for larger projects. The Guide for Artificial Intelligence Ethical Requirements Elicitation[83] consists of 25 cards which are used by the development team to answer questions related to ethical principles. The answers are used to create ethical requirements in the form of ethical user stories which are included in sprint backlogs. ECCOLA [43] consists of 21 cards which are divided into eight themes and with questions to be answered by the development team.

4.2 Design

4.2.1 Multi-Level Co-architecting. Compared with traditional software, the architecture of AI systems is more complex due to different levels of integration. On the one hand, AI models are developed by data scientists/engineers via an AI model pipeline. The AI model pipeline usually is composed of a sequence of automatic steps including data collection, data cleaning, feature engineering, model training, and model evaluation. These steps can be viewed as software components for producing AI models from a software architecture perspective. On the other hand, the produced AI models cannot work alone and need to be integrated into software systems that are to be deployed in the real world [82]. The decisions made by the AI model need to be executed as actions via other software components. The architecture of an AI ecosystem consists of three layers: AI software supply chain, AI system, and operation infrastructure. The focus of the AI software supply chain layer is about developing and managing AI and non-AI components [69], including AI model pipeline components, deployment components, co-versioning components, provenance tracking components, and credential management components, among others. The AI system layer comprises AI components that embed AI models and non-AI components that use the outputs of AI components for overall system functionalities [65]. The operation infrastructure layer is mainly about monitoring and feedback components. Multi-level co-architecting is required to ensure the seamless integration of different components, including co-architecting AI components and non-AI components and co-architecting of different AI model pipeline components. Multi-level co-architecting allows both system- and model-level requirements to be considered in design decision making.

4.2.2 Envisioning Card. AI ethics principles, including the human-centered values principles, are too high level for developers who often lack the technical means to assure human values and ethics. Envisioning cards [17, 115] are designed to help the development team operationalize human values during design processes of AI systems. The design of envisioning cards is based on four envisioning criteria, including stakeholder, time, value, and pervasiveness. The stakeholder criterion helps the development team takes into account the effects of an AI system on both direct stakeholders and indirect stakeholders. The time criterion emphasizes the long-term implication of AI systems on human, society, and environments. The value criterion guides the development team to consider the impact of AI systems on human values. The pervasiveness criterion discusses the challenges encountered if an AI system is widely adopted in terms of geography, culture, demographics, and so on. The adoption of envisioning cards comes at a relatively low cost, in terms of both money and time. However, envisioning cards are hard to scale when the number of participants is large or the AI systems are complex.

4.2.3 Design Modelling for Ethics. To reduce ethical risks, AI ethics principles need to be adhered to during the design process. Design modeling methods can be extended and used to support the modeling of AI components and the ethical aspects, including using UML to describe the architecture of AI systems and represent their ethical aspects [114], designing formal models taking

[83] https://josesiqueira.github.io/RE4AIEthicalGuide/index.html

into account human values [36], using ontologies to model the AI system artifacts for accountability [10, 85], establishing RAI knowledge bases for making design decisions considering ethical concerns [101], and using logic programming to implement ethical principles [8]. UML is an option to describe the AI systems and represent their ethical aspects [114]. The UML extension could be a declarative graphic notation for AI system architecture. Additional stereotypes/metamodel elements can be added for RAI-by-design reference architecture (e.g., to describe AI pipeline components). Use case diagrams can help define the stakeholders and explain the functions they use, which are valuable for achieving accountability. State diagrams are useful to analyze the system states and identify the states that may cause ethical failures. Design patterns like the AI mode switcher can take effect to change the state of an AI system to a more human-controlled state. Sequence diagrams describe the human-AI interactions to ensure all the required explanations are provided. Using design modeling methods is helpful to capture and analyze ethical principles in design. One disadvantage when using modeling languages is the time to create and manage the models. In addition, the modeling languages do not scale up for large and complex systems.

4.2.4 System-Level Ethical Simulation. To avoid ethical disasters and gain public trust, it is necessary to model the real-world situations of AI systems without ethical risk. System-level simulation (e.g., [29, 30, 96, 106]) is a cost-effective way to imitate real-world situations and assess the behaviors of AI systems before deploying AI systems in the real world. A simulation model needs to be built to mimic the possible behaviors and decisions of the AI system and assess the ethical impacts. The assessment results can be sent to the development team or potential users before the AI systems are deployed in the real world. System-level simulation can predict potential ethical risks and avoid serious ethical disasters before deploying AI systems in the real world. However, the simulation model cannot represent all the behaviors and ethical impacts of AI systems in the real world. The accuracy of assessment results is limited by the quality of the simulation model.

4.2.5 Human-Centered Interface Design for Explainable Artificial Intelligence. End users often do not understand how decisions are made by AI systems and are not aware of the capabilities or limitations of the AI systems. The missing explainability may lead to a lack of trust and has been identified as one of the most urgent challenges of RAI to be addressed. **Explainable Artificial Intelligence (XAI)** can be viewed as a human-AI interaction problem and achieved by effective human-centered interface design. Checklists (e.g., a question bank) are often used to help design the explainable user interfaces [62, 66, 67] and understand the user needs, choices of XAI techniques, and XAI design factors [67]. For example, the checklist questions could consider the following aspects [66] for different stakeholders: input, output, how, performance (can be extended to ethical performance), why and why not, what if, and so on. The design of conversational interfaces can be experimented via a Wizard of Oz study [56] in which users interact with a system that they believe to be autonomous but is actually being operated by a hidden human, called the *Wizard*. The conversation data is collected and analyzed to understand requirements for a self-explanatory conversational interface. There can be several ways to increase human trust in AI systems through a human-centered user interface, including anthropomorphism [72] and proactive informing (e.g., capability/limitation of AI systems, ethics credentials, explanations of decisions/behaviors, potential outcomes, data use information).

4.3 Implementation

4.3.1 RAI Governance of APIs. APIs allow developers to solve problems more efficiently and can effectively reduce the development cost and time of AI systems. However, there may be ethical quality issues with APIs (e.g., data privacy breaches or fairness issues). Ethical compliance checking for APIs is needed to detect if any ethics violation exists [50]. A knowledge-driven approach can be adopted to detect ethics issues through ethical knowledge graphs. Ethical knowledge graphs make meaningful entities and concepts, and their relationships in development of AI systems. With the ethical knowledge graph, the rich semantic relationships between entities are explicit and traceable across heterogeneous high-level documents and various AI systems artifacts. Ethical knowledge graphs can be built based on the ethical principles and guidelines (e.g., a privacy knowledge graph based on GDPR [35, 90]) and technical documents (e.g., API documentation) to support the ethical compliance checking for APIs.

4.3.2 RAI Governance via APIs. Some AI systems may provide high-risk capabilities, which can be used or modified to implement harmful tasks. To avoid harmful dual uses in AI systems [80], developers should carefully design how their AI systems can be directly used and indirectly used (i.e., potential ways their systems can be adapted). Developers must restrict the way AI systems are used and preventing the users from getting around of restrictions by unauthorized reverse engineering or modification to the system design. Rather than fully opening the access to AI systems by allowing AI systems to run locally, developers could provide AI services on the cloud and control the interactions with the AI services via APIs [103]. For example, OpenAI's language model GPT-3 can be only integrated with AI systems through an API by approved users.[84] Google Vision AI limits its facial recognition feature to a few celebrities through API.[85]

4.3.3 Ethical Construction with Reuse. Building AI systems from scratch can be very complex and time consuming. Very big companies usually have massive AI investments and large volumes of data to compete in the market, whereas smaller companies may only have a couple of data scientists and can hardly keep up with larger companies. To speed up the development and reduce cost, it is highly desirable and valuable to reuse the AI artifacts (i.e., AI components and/or AI pipeline artifacts) across different applications. However, there might be ethical quality issues with the reused AI artifacts, which requires further assurance mechanisms. Ethical construction with reuse means to develop RAI systems with the use of existing AI artifacts that are compliant with AI ethics principles, such as from an organizational repository or an open source platform. A marketplace can be built up to trade the reusable AI artifacts, including component code, models, and datasets. Blockchain can be adopted to design an immutable and transparent marketplace enabling auction-based trading for AI artifacts and material assets (e.g., cloud resources) [107]. Ethics credentials might be required to be attached to the traded AI artifacts. In addition, tooling support might be needed, such as model migration tool pytorch2keras,[86] and glue code for compatibility.[87] Low/no code tools can also help to achieve ethical construction with reuse.

4.4 Testing

4.4.1 Ethical Acceptance Testing. Since the AI ethical principles are very high level, they need to be captured through ethical requirements, which can be viewed as the agreed commitments by the development team and customers. Ethical acceptance testing (e.g., bias testing) is designed to detect the ethics-related design flaws and verify the ethical requirements (e.g., whether the data pipeline has appropriate privacy control, fairness testing for training/validation data) [3, 20, 123]. In an agile process, the ethical requirements can be framed as ethical user stories and associated with ethical acceptance tests. The ethical acceptance tests are a contract between the customer and

[84] https://openai.com/api/

[85] https://cloud.google.com/vision

[86] https://github.com/gmalivenko/pytorch2keras

[87] https://insights.sei.cmu.edu/blog/software-engineering-for-machine-learning-characterizing-and-detecting-mismatch-in-machine-learning-systems/

development team. The behavior of the AI system should be quantified by the acceptance tests, and the acceptance criteria for each of the ethical principles should be defined in a testable way. The history of ethical acceptance testing should be recorded and tracked, such as how and by whom the ethical issues were fixed. A testing leader may be appointed to lead the ethical acceptance testing for each ethics principle. For example, when bias is detected at runtime, the monitoring reports are returned to the bias testing leader [28, 104]. Ethical acceptance tests capture the ethical requirements and measure how well the AI system meets ethical requirements, but they may need to be amended frequently as ethical requirements change.

4.4.2 Ethical Assessment for Test Cases. The ethical quality assurance for AI systems is heavily dependent on ethical acceptance testing, which is aimed at detecting and solving ethical issues in the AI system. A collection of test cases with expected results should be generated [83] and maintained to detect possible ethical failures in a variety of extreme situations [42]. However, there might be ethical issues within the test cases. For example, the test data may introduce fairness or privacy issues [77]. Preparing quality test cases is an integral part of ethical acceptance testing. A test case usually is composed of the ID, description, preconditions, test steps, test data, expected results, actual results, status, creator name, creation date, executor name, and execution date. All the test cases for verification and validation should pass the ethics assessment. This includes ethical risk assessment for test steps and test data. The creation and execution information are essential to track the accountability of ethical issues with test cases. Ethical assessment for test cases improves the ethical quality of the development process of AI systems, but new test cases need to be continually added and assessed when a new ethical requirement is added or the operation context changes.

4.5 Operation

4.5.1 Continuous Deployment for RAI. AI systems may frequently evolve due to their data dependency. When ethical performance degradation occurs over time, AI models need to be retrained with new data or features and reintegrated into AI components. The non-AI component may also need to be upgraded to meet new requirements or changing context. New versions of AI systems need to be frequently and continuously deployed into production environments. However, AI systems involve a higher degree of uncertainty and risks associated with the autonomy of the AI systems. Thus, there is a strong desire for various deployment strategies to support continuous deployment [75, 128]. There are various deployment strategies for AI systems. Phased deployment means deploying AI systems for a subset group of users initially to reduce ethical risk [47]. The new version of AI systems rolls out incrementally and serves alongside the old version. Phased deployment can also be about automating decisions in phases to better supervise and control automation. This usually depends on the stakes of the situations and the level of confidence that users may have with automatic decisions made by AI systems. Further, A/B testing deployment [58] is a common deployment strategy undertaken in industry, where different versions of the AI model are deployed to production. The models are compared and selected based on their ethical performance. In addition, the existing reliability practices, like redundancy, are also applicable to AI components in an AI system. Multiple AI models work independently to improve the ethical performance of the AI components. Applying various deployment strategies helps to reduce the ethical risk. Users can be quickly redirected to the older version or the other version of AI systems/models. However, it is complex and expensive to adopt different deployment strategies during operations.

4.5.2 Extensible, Adaptive, and Dynamic Ethical Risk Assessment. The current risk-based approach to ethical principles is often a done-once-and-forget type of algorithm-level risk assessment [45, 73, 99, 104, 130] and mitigation for a subset of ethical principles (e.g., privacy or fairness[88]) at a particular development step (e.g., Canada's Algorithmic Impact Assessment Tool[89]), which is not sufficient for the highly uncertain and continual learning AI systems. In addition, the context of AI systems varies with the application domains, organizations, culture, and regions. It is essential to perform continuous risk assessment and mitigation of RAI systems [25, 111]. The ethical risk assessment framework can be built with guided extension points for different contexts (e.g., culture context). The risk mitigation can be designed from three aspects: reducing frequency occurrence, consequence size, and consequence response. Extensible, adaptive, and dynamic risk assessment can effectively ensure that an AI system adheres to AI ethics principles throughout the whole lifecycle, but it might be hard to measure some of the ethical principles, such as human-centered values.

4.5.3 Multi-Level Co-versioning. AI systems involve two levels of relationships and dependencies across various AI artifacts, including the supply chain level and the system level. At the system level, there are multiple versions of AI components and non-AI components. At the supply chain level, there are different versions of data, model, code, and configuration, which are used to produce different versions of AI components [65]. At the system level, the AI components that embed AI models are integrated into AI systems and interact with non-AI components. However, the retraining of AI models introduces new versions of data, code, and configuration parameters. If federated learning is adopted, for each round of training, a global model is ensembled based on local models sent from participating clients [70]. It is important to capture all of these dependencies during the development process. Multi-level co-versioning provides end-to-end traceability and accountability throughout the whole lifecycle of AI systems, but the collection and documentation of co-versioning information incur additional development cost. There have been many version control tools in industry focusing on supply chain level co-versioning, such as MLflow Model Registry on Databricks[90] and the Amazon provenance tool,[91] and Data Version Control (DVC).[92]

5 PRODUCT PATTERNS

This section provides system-level guidance on how to design the architecture of RAI systems. We present a collection of product patterns (i.e., design patterns) for building RAI-by-design into AI systems (Figure 6). Broadly, an AI system is composed of three layers: (1) the supply chain layer that generates the software components which compose the AI system, (2) the system layer which is the deployed AI system, and (3) the operation infrastructure layer that provides auxiliary functions to the AI system. Figure 7 presents the identified products patterns for each of the three layers. Those product patterns can be embedded into the AI ecosystems as product features. Figure 7 illustrates a state diagram of a provisioned AI system and highlights the patterns associating with relevant states or transitions, which show when the product patterns could take effect.

5.1 Supply Chain Patterns

5.1.1 Bill of Materials Registry. The bill of materials registry [12, 88] can be designed to keep a formal machine-readable record of the supply chain details of the components used in building

[88] https://www.nist.gov/artificial-intelligence/proposal-identifying-and-managing-bias-artificial-intelligence-sp-1270
[89] https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html
[90] https://docs.databricks.com/applications/mlflow/model-registry.html
[91] https://www.amazon.science/publications/automatically-tracking-metadata-and-provenance-of-machine-learning-experiments
[92] https://dvc.org/

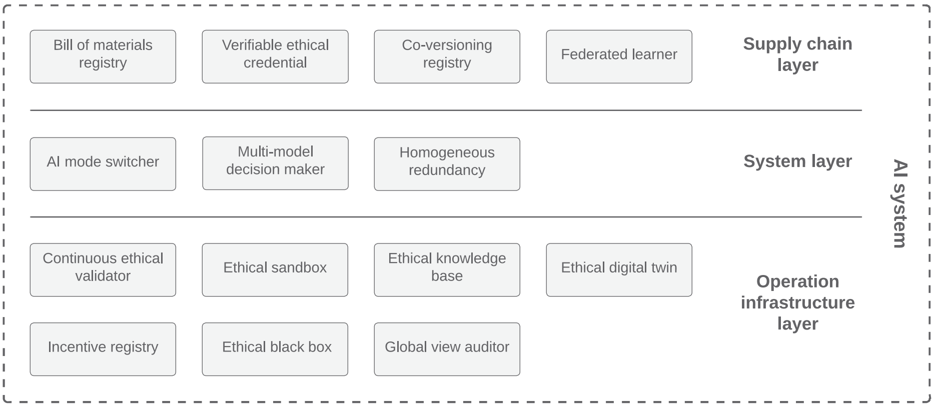


Fig. 6. Product patterns for RAI-by-design architecture of an AI system.

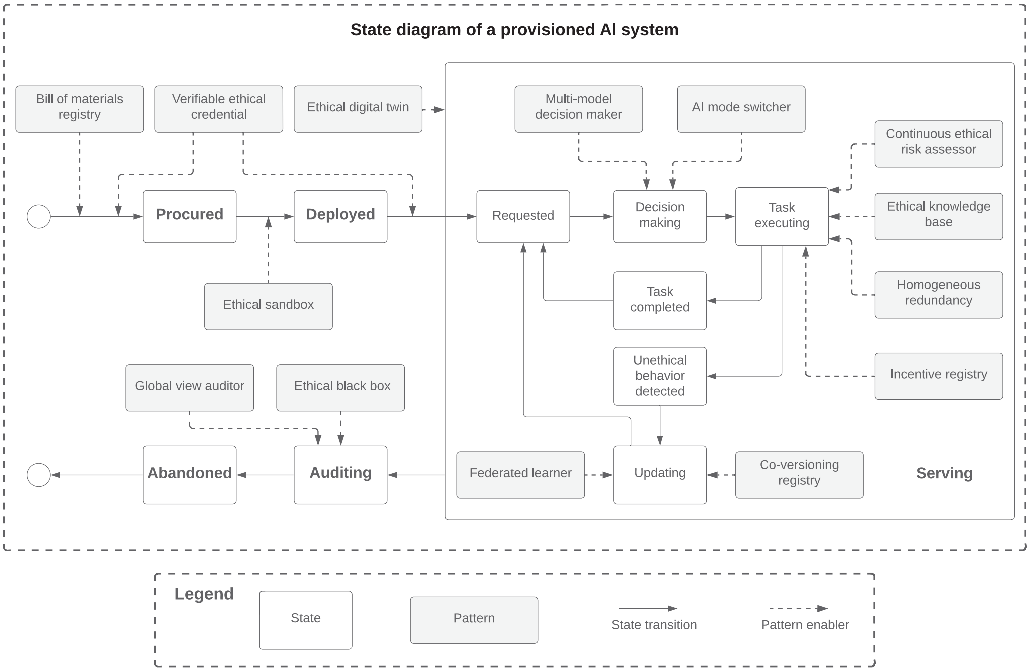


Fig. 7. Product patterns for RAI-by-design.

an AI system, including component name, version, supplier, dependency relationship, author, and timestamp. In addition to supply chain details of the components, context documents (like model cards [81] for reporting AI models, and datasheets for the datasets [41] used to train AI models) can be integrated to the bill of materials registry.

The main purpose of the bill of materials registry is to provide traceability and transparency into the components within AI systems so that ethical issues can be tracked and addressed [11]. Some platforms manage the bill of materials registry, such as OpenBOM,[93] Codenotary,[94] and Snorkel

[93] https://www.openbom.com/
[94] https://codenotary.com/

Flow.[95] An immutable data infrastructure can store the bill of materials to enable integrity. For example, the manufacturers of autonomous vehicles could maintain a material registry contract on blockchain to track their components' supply chain information (e.g., the version and supplier of the third-party navigation component). Stakeholders can access the supply chain details of each component of interest in AI systems via the bill of materials registry. As AI systems evolve over time, the bill of materials may need to be updated frequently. The cost of managing the bill of materials of all the components depends on the complexity of the AI system.

5.1.2 Verifiable Ethical Credential. Verifiable ethical credentials can be used as evidence of ethical compliance for AI systems, components, models, developers, operators,[96] users, organizations, and development processes [21, 72, 92]. Verifiable credentials are data that could be cryptographically verified and be presented with strong proofs [23]. Publicly accessible data infrastructure needs to be built to support the generation and verification of the ethical credentials on a neutral platform. Before using AI systems, users may verify the systems' ethical credential to check if the systems are compliant with AI ethics principles or regulations [21]. However, the users may be required to provide the ethical credentials to use and operate the AI systems (e.g., to ensure the flight safety of drones). The verifiable ethical credential helps increase user trust toward an AI system through conferring the trust that the user has with the authority that issues the credential to AI systems, organizations that develop AI systems, and operators who operate AI systems. Such transitive trust relationship is critical to the efficient functioning of the AI system. With an ethical credential, an AI system could provide proof of compliance as an incentive for the users to use the AI system, thus increasing AI adoption. An ethical credential may be forged, which makes the verification of authenticity of the ethical credentials become challenging. Blockchain could be adopted to build the credential infrastructure to ensure data integrity. For example, SecureKey[97] is a blockchain-based infrastructure for ID management with support of a verifiable credential.

5.1.3 Co-versioning Registry. Compared with traditional software, AI systems involve different levels of dependencies and may evolve more frequently due to their data-dependent behaviors. From the viewpoint of the AI system, it is important to know the version of the AI component integrated into the system. From the viewpoint of the AI component, it is important to know what datasets and parameters were used to train the AI model and what data was used to evaluate the AI model. Co-versioning of the components or AI artifacts of AI systems provides end-to-end provenance guarantees across the entire lifecycle of AI systems. The co-versioning registry can track the co-evolution of components or AI artifacts [65, 70]. There are different levels of co-versioning: co-versioning of AI components and non-AI components, co-versioning of the artifacts within the AI components (i.e., co-versioning of data, model, code, and configurations, and co-versioning of local models and global models in federated learning). Co-versioning enables effective maintenance and evolution of AI components because the deployed model or code can be traced to the exact set of artifacts, parameters, and metadata that were used to develop the model and code. The MLflow Model Registry[98] is a model repository and set of APIs that enable management of the full lifecycle of MLflow Models, including model lineage and versioning.

5.1.4 Federated Learner. Despite the widely deployed mobile or IoT devices generating massive amounts of data, lack of training data is still a challenge for AI systems given the increasing concern in data privacy. The federated learner trains an AI model across multiple edge devices or

[95] https://snorkel.ai/
[96] https://certnexus.com/certification/ceet/
[97] https://securekey.com/
[98] https://docs.databricks.com/applications/mlflow/model-registry.html

servers with local data samples. The federated learner [15, 18, 68–70, 112, 113, 117] preserves the data privacy by training models locally on the client devices and formulating a global model on a central server based on the local model updates (e.g., train the visual perception model locally in each vehicle). Decentralized learning is a variant of federated learning, which could use blockchain to remove the single point of failure and coordinate the learning process in a fully decentralized way [120]. TensorFlow Federated[99] is an open source framework for machine learning on decentralized data sources. FATE[100] is an open source project that supports the federated AI ecosystem.

5.2 System Patterns

5.2.1 AI Mode Switcher. When to use AI at decision-making points can be a major architectural design decision when designing an AI system. Adding an AI mode switcher to the AI system offers users efficient invocation and dismissal mechanisms for activating and deactivating the AI component whenever needed, thus defer the architectural decision to the execution time that is decided by the end user or the operator of the AI system. The AI mode switcher is like a kill switch of an AI system that could immediately shut down the AI component and thus stop its negative effects [78, 89, 116] (e.g., turning off the automated driving system and disconnecting it from the internet). The decisions made by the AI component can be executed automatically or reviewed by a human expert before being executed in critical situations. The human expert serves to approve or override the decisions (e.g., skipping the path generated by the navigation system). Human intervention can also happen after acting on the AI decision through the fallback mechanism that reverses the system back to the state before executing the AI decision. A built-in guard can be used to ensure that the AI component is only activated within the pre-defined conditions (e.g., domain of use, boundaries of competence). The end users or the operators can ask questions or report complaints/failures/near misses through a recourse channel after observing a bad decision from the AI component. Tesla Autopilot[101] has multiple driver assistance features that can be enabled or disabled during the driving. Users maintain control of the vehicles and can override the operations by these features at runtime. The Baidu autonomous mini-bus Robobus[102] requires a staff in the seat to supervise the self-driving operations, and the bus can be switched to manual driving mode by braking.

5.2.2 Multi-Model Decision Maker. In the reliability community of software systems, traditional architecture-based software reliability is based on a software component. The existing reliability practices, like redundancy, are also applicable to AI components in an AI system. In addition, a reasonable combination of multiple AI models that are normally work independently could improve performance (e.g., accuracy) of the AI component. The multi-model decision maker employs different models to perform the same task or enable a single decision (e.g., deploying different algorithms for visual perception). It improves the reliability by deploying different models under different contexts (e.g., different geo-location regions) and enabling fault tolerance by cross validating ethical requirements for a single decision [24, 84]. Different consensus protocols could be defined to make the final decision—for example, taking the majority decision. Another strategy is to only accept the same results from the employed models. In addition, the end user or the operator could step in to review the output from the multiple models and make a final decision based on human expertise. Scikit-learn[103] is a Python package that supports using multiple learning algorithms to obtain

[99] https://www.tensorflow.org/federated
[100] https://fate.fedai.org/
[101] https://www.tesla.com/autopilot
[102] https://apollo.auto/minibus/
[103] https://github.com/scikit-learn/scikit-learn

better performance through ensemble learning. The AWS Fraud Detection Using Machine Learning solution trains an unsupervised anomaly detection model in addition to a supervised model, to augment the prediction results.[104] IBM Watson Natural Language Understanding uses an ensemble learning framework to include predictions from multiple emotion detection models.[105]

5.2.3 Homogeneous Redundancy. *N*-version programming is a software design pattern to ensure fault tolerance of software [61]. Similarly, deploying multiple redundant and identical AI components (e.g., two brake control components) can be a solution to tolerate the individual AI component with high uncertainty that may make unethical decisions or the individual adversary hardware component that produces malicious data or behaves unethically [84]. A cross check can be conducted for the outputs provided by multiple components of a single type. The results are accepted only as there is a consensus among the redundant components. The results that are not accepted automatically according to a consensus protocol can be further reviewed by the end user or the operator of the AI system. Waymo[106] contains multiple redundant components at various levels, including redundant braking, steering, and inertial measurement systems for vehicle positioning.

5.3 Operation Infrastructure Patterns

5.3.1 Continuous Ethical Validator. AI components of an AI system often require continual learning based on new data collected during operation of the AI system. The continuous ethical validator deployed in an AI system continuously monitors and validates the outcomes of AI components (e.g., the path recommended by the navigation system) against the ethical requirements [58, 111]. The outcomes of AI systems are about whether the AI system provides the intended benefits and behaves appropriately given the situation. The time and frequency of validation can be configured. Version-based feedback and rebuild alert are sent when the pre-defined conditions regarding the ethical requirement are met. AWS SageMaker Model Monitor[107] continuously monitors the bias drift of the AI models in production. Qualdo[108] is an AI monitoring solution that monitors data quality and model drift. Azure Machine Learning[109] uses Azure Monitor to create monitoring data. Azure Monitor is a full-stack monitoring service.

5.3.2 Ethical Sandbox. Given that AI systems are of high stake, it is risky to run the entire system in the same execution environment. Ethical sandbox can be applied to isolate an AI component from other AI components and non-AI components by running the AI component separately in a safe environment [63] (e.g., sandboxing the unverified visual perception component). Thus, the AI component could execute without affecting other components and the output of the AI system. The ethical sandbox is an emulated environment with no access to the rest of the AI system. An emulation environment duplicates all the hardware and software functionality of an AI system. Thus, developers could run an AI component safely to determine how it works and whether it is responsible before widely deploying the AI component. Maximal tolerable probability of violating the ethical requirements should be defined as the ethical margin for the sandbox. A watchdog can be used to limit the execution time of the AI component to reduce the ethical risk (e.g., only activating the visual perception component for 5 minutes on the bridges

[104] https://aws.amazon.com/solutions/implementations/fraud-detection-using-machine-learning/
[105] https://www.ibm.com/au-en/cloud/watson-natural-language-understanding
[106] https://waymo.com/
[107] https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor.html
[108] https://www.qualdo.ai/monitor-ml-model-performance-monitoring/
[109] https://docs.microsoft.com/en-us/azure/machine-learning/monitor-azure-machine-learning

built especially for autonomous vehicles). Fastcase AI Sandbox[110] provides a secure platform for users to upload dataset and do data analysis in a safe environment. AI Sandbox[111] provides an AI execution and RESTful interface that could be used by modern programming languages.

5.3.3 *Ethical Knowledge Base.*

The ecosystem of AI systems involves broad ethical knowledge, such as AI ethics principles, regulations, and guidelines. Such ethical knowledge is scattered and is usually implicit or abstract to end users or even developers and data scientists who primarily are without a legal background and focus more on the technical aspects of AI systems. An ethical knowledge base, such as a knowledge graph, makes meaningful entities and concepts, and their relationships in design, implementation, deployment, and operation of AI systems [32, 85, 101]. With the ethical knowledge base, the rich semantic relationships between entities are explicit and traceable across heterogeneous high-level documents on one hand and different artifacts across the AI system lifecycle on the other hand. Thus, ethical requirements of the AI system can be systematically accessed and analyzed using the ethical knowledge base. Awesome AI Guidelines[112] aims to provide a mapping between ecosystem of guidelines, principles, codes of ethics, standards, and regulation around AI. The RAI community portal[113] is provided by AI Global, which is an evolving repository of reports, standards, models, government policies, datasets, and open source software to inform and support RAI development. Responsible AI Knowledge-base[114] is a knowledge base of different areas using and developing AI in a responsible way.

5.3.4 *Ethical Digital Twin.*

Simulation is designed to imitate a real-world situation. Before running an AI system in the real world, it is important to perform system-level simulation through an ethical digital twin running on a simulation infrastructure to understand the behaviors of the AI system and assess ethical risks in a cost-effective way. Digital twin [102] was introduced by NASA as a digital representation of a real system used in lab-testing activities. The digital twin of an AI system could be used to represent the behaviors of the AI system and forecast change impacts. The ethical digital twin can also be used during operation of the AI system to assess the system's runtime behaviors and decisions based on the simulation model using the real-time data. The assessment results can be sent back to alert the system or user before the unethical behavior or decision takes effect [29]. Vehicle manufacturers can use the ethical digital twin to explore the limits of autonomous vehicles based on the collected real-time data, such as NVIDIA DRIVE Sim[115] and rfPro.[116]

5.3.5 *Incentive Registry.*

Incentive mechanisms are effective treatments in motivating AI systems and encouraging the stakeholders involved in the AI system ecosystem to execute tasks in a responsible manner. An incentive registry records the rewards that correspond to the AI system's ethical behavior and outcome of decisions [121, 125] (e.g., rewards for path planning without ethical risks). There are various ways to formulate the incentive mechanism, such as using reinforcement learning or building the incentive mechanism on a publicly accessible data infrastructure like blockchain [125]. Traditional incentive mechanisms for human participants include reputation based and payment based. However, it is challenging to formulate the form of rewards in the context of RAI, as the ethical impact of AI systems' decisions and behaviors

[110] https://www.fastcase.com/sandbox/
[111] https://aisandbox.dev/
[112] https://github.com/EthicalML/awesome-artificial-intelligence-guidelines
[113] https://portal.ai-global.org/
[114] https://github.com/alexandrainst/responsible-ai
[115] https://developer.nvidia.com/drive/drive-sim
[116] https://www.rfpro.com/

might hardly be measured for some of the ethical principles (e.g., human values). Furthermore, the incentive mechanism needs to be agreed on by all stakeholders, who may have different views on the ethical impact. In addition, there may be tradeoffs between different principles, which makes the design harder. The Open Science Rewards and Incentives Registry[117] incentivizes the development of an academic career structure that fosters outputs, practices, and behaviors to maximize contributions to a shared research knowledge system. FLoBC[118] is a tool for federated learning over blockchain that utilizes a reward/punishment policy to incentivize legitimate training, and to punish and hinder malicious trainers.

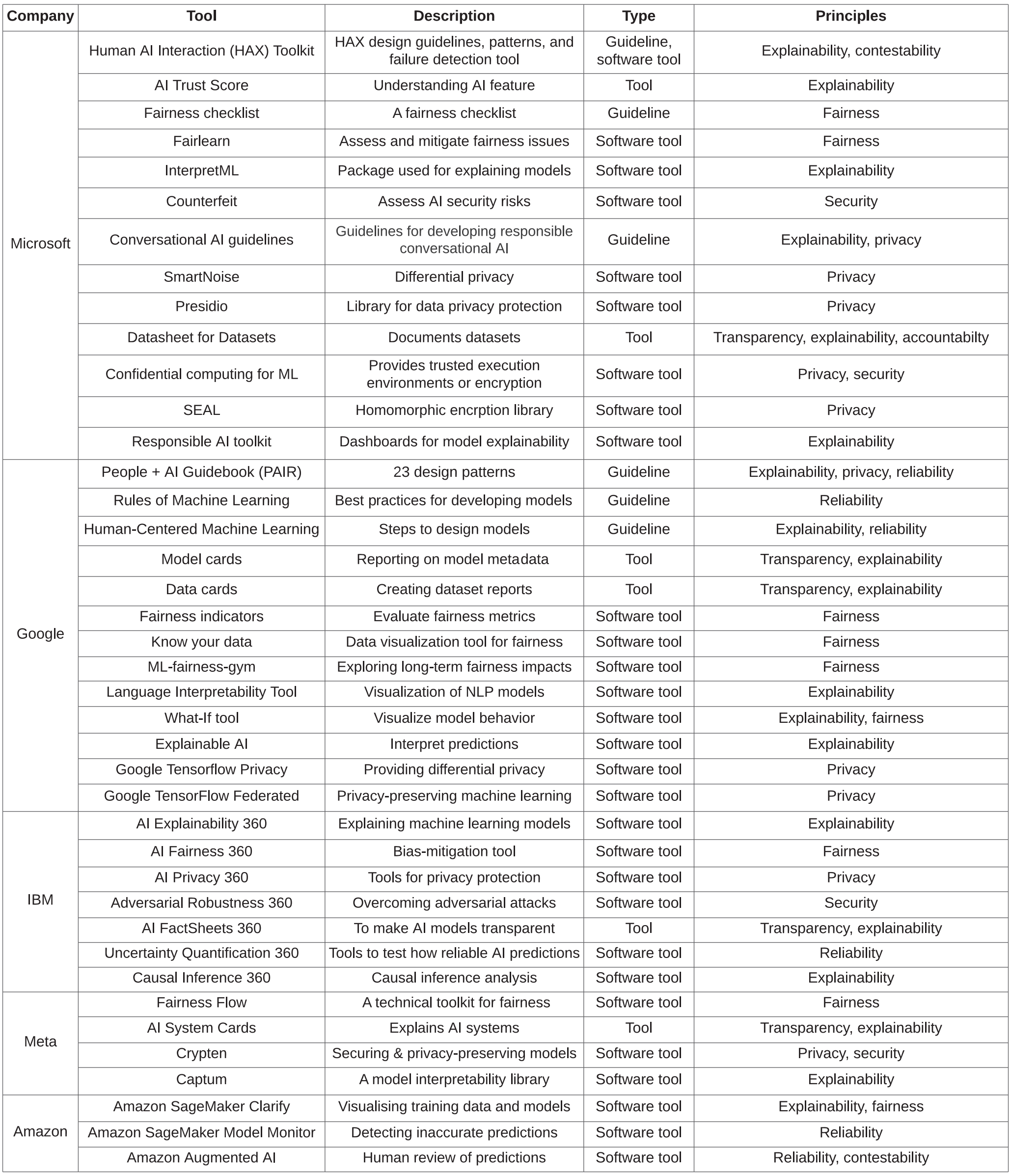
5.3.6 Ethical Black Box. The black box was introduced initially for aircraft several decades ago for recording critical flight data. The intention of adding a black box to aircrafts is to collect evidence of the actions of system and the surrounding context information for analysis after near misses and failures. The near misses and failures are specific to the use cases. Although the primary usage of a black box is accident investigation, black boxes are useful for other purposes. Data collection and the analysis could support improvement of the system. The purpose of embedding an ethical black box in an AI system is to investigate why and how an AI system caused an accident or a near miss. The ethical black box continuously records sensor data, internal status data, decisions, behaviors (both system and operator), and effects [33, 34, 122]. For example, an ethical black box could be built into an automated driving system to record the behaviors of the system and driver and their effects [34]. All of these data need to be kept as evidence with the timestamp and location data. Designing the ethical black box is challenging, as the ethical metrics need to be identified for data collection. In addition, design decisions need to be made on what data should be recorded and where the data should be stored (e.g., using a blockchain-based immutable log or a cloud-based data storage). RoBoTIPS[119] aims to develop an ethical black box for social robots, to enable the explainability of their behavior.

5.3.7 Global View Auditor. When an accident happens, there might be more than one AI system or multiple AI components within one AI system involved (e.g., multiple autonomous vehicles in an accident). The data collected from each involved AI system/component might conflict with each other since the individual AI system/component may have their own perception. The global-view auditor is a component that collects information from multiple AI components/AI systems and processes the information to identify discrepancies among the information collected [79]. Based on the result, the global-view auditor may alert the AI system/component to a wrong perception, thus avoiding negative impacts or identifing liability when negative events occur. This pattern can be also used to improve the decision making of an AI system by taking the knowledge from other systems. For example, an autonomous vehicle may increase its visibility using the perceptions of others to make better decisions at runtime. The global-view auditor enables accountability that covers different perceptions of AI components/systems that are involved and redresses the conflicting information collected from multiple AI components/systems.

6 RELATED WORK

The challenge of RAI has attracted significant attention in both industry and academia. To achieve RAI, there have been nearly 100 high-level AI ethics principles and guidelines issued by governments, organizations, and companies [57]. Some degree of consensus around AI ethics principles has been achieved [37]. A principle-based approach allows technology-independent

[117] https://opensciencereqistry.org/
[118] https://github.com/Oschart/FLoBC
[119] https://www.robotips.co.uk/



Company	Tool	Description	Type	Principles
Microsoft	Human AI Interaction (HAX) Toolkit	HAX design guidelines, patterns, and failure detection tool	Guideline, software tool	Explainability, contestability
	AI Trust Score	Understanding AI feature	Tool	Explainability
	Fairness checklist	A fairness checklist	Guideline	Fairness
	Fairlearn	Assess and mitigate fairness issues	Software tool	Fairness
	InterpretML	Package used for explaining models	Software tool	Explainability
	Counterfeit	Assess AI security risks	Software tool	Security
	Conversational AI guidelines	Guidelines for developing responsible conversational AI	Guideline	Explainability, privacy
	SmartNoise	Differential privacy	Software tool	Privacy
	Presidio	Library for data privacy protection	Software tool	Privacy
	Datasheet for Datasets	Documents datasets	Tool	Transparency, explainability, accountabilty
	Confidential computing for ML	Provides trusted execution environments or encryption	Software tool	Privacy, security
	SEAL	Homomorphic encrption library	Software tool	Privacy
	Responsible AI toolkit	Dashboards for model explainability	Software tool	Explainability
Google	People + AI Guidebook (PAIR)	23 design patterns	Guideline	Explainability, privacy, reliability
	Rules of Machine Learning	Best practices for developing models	Guideline	Reliability
	Human-Centered Machine Learning	Steps to design models	Guideline	Explainability, reliability
	Model cards	Reporting on model metadata	Tool	Transparency, explainability
	Data cards	Creating dataset reports	Tool	Transparency, explainability
	Fairness indicators	Evaluate fairness metrics	Software tool	Fairness
	Know your data	Data visualization tool for fairness	Software tool	Fairness
	ML-fairness-gym	Exploring long-term fairness impacts	Software tool	Fairness
	Language Interpretability Tool	Visualization of NLP models	Software tool	Explainability
	What-If tool	Visualize model behavior	Software tool	Explainability, fairness
	Explainable AI	Interpret predictions	Software tool	Explainability
	Google Tensorflow Privacy	Providing differential privacy	Software tool	Privacy
	Google TensorFlow Federated	Privacy-preserving machine learning	Software tool	Privacy
IBM	AI Explainability 360	Explaining machine learning models	Software tool	Explainability
	AI Fairness 360	Bias-mitigation tool	Software tool	Fairness
	AI Privacy 360	Tools for privacy protection	Software tool	Privacy
	Adversarial Robustness 360	Overcoming adversarial attacks	Software tool	Security
	AI FactSheets 360	To make AI models transparent	Tool	Transparency, explainability
	Uncertainty Quantification 360	Tools to test how reliable AI predictions	Software tool	Reliability
	Causal Inference 360	Causal inference analysis	Software tool	Explainability
Meta	Fairness Flow	A technical toolkit for fairness	Software tool	Fairness
	AI System Cards	Explains AI systems	Tool	Transparency, explainability
	Crypten	Securing & privacy-preserving models	Software tool	Privacy, security
	Captum	A model interpretability library	Software tool	Explainability
Amazon	Amazon SageMaker Clarify	Visualising training data and models	Software tool	Explainability, fairness
	Amazon SageMaker Model Monitor	Detecting inaccurate predictions	Software tool	Reliability
	Amazon Augmented AI	Human review of predictions	Software tool	Reliability, contestability

Fig. 8. Top five major industry players on RAI according to the number of tools.

operationalization of RAI. However, these principles are very abstract and high level for stakeholders of AI systems to use in practice.

Significant efforts have been put on algorithm-level solutions which mainly focus on a subset of principles. Figure 8 lists the top five major industry players on RAI according to their number of RAI tools based on the results of our MLR study. Most of these tools focus on privacy, security, reliability, safety, fairness, and explainability from an AI model perspective. More work is needed on transparency, accountability, contestability, and human-centered values, and human, societal, and environmental wellbeing, particularly from a system perspective.

Overall, AI ethics principles need to be operationalized in the form of concrete patterns and best practices that are usable by AI developers and other stakeholders to build up RAI systems. Some ad-hoc sets of guidebooks, question banks, checklists, and templates have started to appear. Microsoft's Human AI Interaction (HAX) Toolkit provides a set of HAX guidelines and patterns.[120] However, those guidelines and patterns only focus on interaction design and do not provide any guidance on development and governance. Google's PAIR (People + AI Research) guidebook summarizes 23 design patterns[121] which mainly address some of the AI ethics principles for AI models, including explainability, privacy, and reliability. Process and governance guidelines are not discussed in Google's PAIR. Although OECD provides a framework of tools for trustworthy AI [87], the framework largely contains categorized but disjointed software tools, lacking process-related linkages. Thus, a systematic and operationalized guidance for AI system stakeholders is required throughout the entire lifecycle of AI systems.

There have been a few survey papers on operationalizing RAI [7, 99, 109]. However, the findings and insights in these papers are still around principles and do not provide concrete and actionable guidelines for stakeholders to use in practice. Our previous roadmap paper [71] discusses the current state and identifies the critical research challenges in the area of software engineering for RAI based on an initial SLR. This pattern catalogue paper is built on top of the published roadmap and provides a comprehensive list of concrete patterns from multi-level governance patterns to process and product patterns based on the results of an MLR. In the RAI Pattern Catalogue, we present structured knowledge about the patterns, including context, problem, solution, benefits, drawbacks, and known uses. The full version of the RAI Pattern Catalogue is available online.[4]

7 THREATS TO VALIDITY

External Validity. First, "responsible artificial intelligence" is loosely defined with many terms that refer to emerging technologies in this area. There is a set of terms currently being used in the community to mean largely the same thing: RAI, AI ethics, ethical AI, trustworthy AI, and trust in AI. This issue has been addressed by including search terms that are being used interchangeably in the search string to ensure that all relevant works were covered. Another issue is that many solutions were initially designed only for addressing one of the AI ethics principles but could be identified as a pattern and extended to implement RAI. To mitigate this threat, we included all AI ethics principles in the search string as supplementary terms.

Internal Validity. To mitigate the threat of not finding all relevant studies, we performed a rigorous search using defined keywords and executed snowballing that allows us to recover the missing studies from the literature. To address the bias, one researcher performed the screening of titles, abstracts, and full texts. The other researcher evaluated a random sample of the selected studies after screening to check the consistency of their inclusion/exclusion decisions.

8 CONCLUSION

To operationalize RAI, this article adopts a pattern-oriented approach and presented a comprehensive RAI Pattern Catalogue that AI system stakeholders can utilize to ensure that the developed AI systems are trustworthy throughout the entire governance and engineering lifecycle, from multi-level governance patterns to concrete process and product patterns. These patterns offer a systematic and actionable system-level guidance with consequence analysis and well-known uses for AI system stakeholders to reference during the governance and development processes. We are currently building a Question Bank and a software tool for AI risk assessment, which will use the RAI

[120] https://www.microsoft.com/en-us/haxtoolkit/ai-guidelines/
[121] https://pair.withgoogle.com/guidebook/patterns

Pattern Catalogue as one of the knowledge sources to recommend mitigation strategies. We also plan to validate the utility, usability, and effectiveness of the pattern catalogue and the supporting tools in industrial projects.

REFERENCES

[1] David Adkins, Bilal Alsallakh, Adeel Cheema, Narine Kokhlikyan, Emily McReynolds, Pushkar Mishra, Chavez Procope, Jeremy Sawruk, Erin Wang, and Polina Zvyagina. 2022. Prescriptive and descriptive approaches to machine-learning transparency. In *CHI'22: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–9.

[2] David Adkins, Bilal Alsallakh, Adeel Cheema, Narine Kokhlikyan, Emily McReynolds, Pushkar Mishra, Chavez Procope, Jeremy Sawruk, Erin Wang, and Polina Zvyagina. 2022. Method cards for prescriptive machine-learning transparency. In *Proceedings of the 2022 IEEE/ACM 1st International Conference on AI Engineering–Software Engineering for AI (CAIN'22)*. IEEE, Los Alamitos, CA, 90–100.

[3] Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. 2019. Black box fairness testing of machine learning models. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 625–635.

[4] Mohit Kumar Ahuja, Mohamed-Bachir Belaid, Pierre Bernabé, Mathieu Collet, Arnaud Gotlieb, Chhagan Lal, Dusica Marijan, Sagar Sen, Aizaz Sharif, and Helge Spieker. 2020. Opening the software engineering toolbox for the assessment of trustworthy AI. *arXiv preprint arXiv:2007.07768* (2020).

[5] Sulaiman Alsheiabni, Yen Cheung, and Chris Messom. 2019. Towards an artificial intelligence maturity model: From science fiction to business facts. In *Proceedings of the Pacific Asia Conference on Information Systems (PACIS'19)*.

[6] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: A case study. In *Proceedings of the 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP'19)*. IEEE, Los Alamitos, CA, 291–300.

[7] Marianna Anagnostou, Olga Karvounidou, Chrysovalantou Katritzidaki, Christina Kechagia, Kyriaki Melidou, Eleni Mpeza, Ioannis Konstantinidis, Eleni Kapantai, Christos Berberidis, Ioannis Magnisalis, and Vassilios Peristersas. 2022. Characteristics and challenges in the industries towards responsible AI: A systematic literature review. *Ethics and Information Technology* 24, 3 (2022), 1–18.

[8] Michael Anderson and Susan Leigh Anderson. 2018. GenEth: A general ethical dilemma analyzer. *Paladyn, Journal of Behavioral Robotics* 9, 1 (2018), 337–357.

[9] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K. Natesan Ramamurthy, D. Reimer, Alexandra Olteanu, David Piorkowski, J. Tsay, and K. R. Varshney. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4-5 (2019), Article 6, 13 pages.

[10] Pelin Ayranci, Phung Lai, Nhathai Phan, Han Hu, Alexander Kolinowski, David Newman, and Deijing Dou. 2022. OnML: An ontology-based approach for interpretable machine learning. *Journal of Combinatorial Optimization* 44 (2022), 770–793.

[11] Iain Barclay, Alun Preece, Ian Taylor, Swapna Krishnakumar Radha, and Jarek Nabrzyski. 2023. Providing assurance and scrutability on shared data and machine learning models with verifiable credentials. *Concurrency and Computation: Practice and Experience* 35, 18 (2023), e6997.

[12] Iain Barclay, Alun Preece, Ian Taylor, and Dinesh Verma. 2019. Towards traceability in data ecosystems using a bill of materials model. In *Proceedings of the International Workshop on Science Gateways*.

[13] Kent Beck and Ward Cunningham. 1987. Using pattern languages for object oriented programs. In *Proceedings of the Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA'87)*.

[14] Richard Benjamins, Alberto Barbado, and Daniel Sierra. 2019. Responsible AI by design in practice. *arXiv preprint arXiv:1909.12838* (2019).

[15] Stefano Bennati and Catholijn M. Jonker. 2017. PriMaL: A privacy-preserving machine learning method for event detection in distributed sensor networks. *arXiv preprint arXiv:1703.07150* (2017).

[16] Adrien Bibal, Michael Lognoul, Alexandre De Streel, and Benoît Frénay. 2021. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law* 29, 2 (2021), 149–169.

[17] Karl-Emil Kjær Bilstrup, Magnus H. Kaspersen, and Marianne Graves Petersen. 2020. Staging reflections on ethical dilemmas in machine learning: A card-based design workshop for high school students. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1211–1222.

[18] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 1175–1191.

[19] Pal Boza and Theodoros Evgeniou. 2021. *Implementing AI Principles: Frameworks, Processes, and Tools.* Working Paper No. 2021/04/DSC/TOM. INSEAD.

[20] Ankur Chattopadhyay, Abdikadar Ali, and Danielle Thaxton. 2021. Assessing the alignment of social robots with trustworthy AI design guidelines: A preliminary research study. In *Proceedings of the 11th ACM Conference on Data and Application Security and Privacy*. 325–327.

[21] Wenjing Chu. 2022. A decentralized approach towards responsible AI in social ecosystems. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 79–89.

[22] Peter Cihon, Moritz J. Kleinaltenkamp, Jonas Schuett, and Seth D. Baum. 2021. AI certification: Advancing ethical practice by reducing information asymmetries. *IEEE Transactions on Technology and Society* 2, 4 (2021), 200–209.

[23] World Wide Web Consortium. 2019. Verifiable Credentials Data Model 1.0: Expressing Verifiable Information on the Web. Retrieved October 14, 2023 from https://www.w3.org/TR/vc-data-model/?#core-data-model

[24] Jian Dai, Shuge Lei, Licong Dong, Xiaona Lin, Huabin Zhang, Desheng Sun, and Kehong Yuan. 2021. More reliable AI solution: Breast ultrasound diagnosis using multi-AI combination. *arXiv preprint arXiv:2101.02639* (2021).

[25] Mathieu d'Aquin, Pinelopi Troullinou, Noel E. O'Connor, Aindrias Cullen, Gráinne Faller, and Louise Holden. 2018. Towards an "ethics by design" methodology for AI research projects. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 54–59.

[26] Virginia Dignum. 2019. Ensuring responsible AI in practice. In *Responsible Artificial Intelligence*. Springer, 93–105.

[27] DISER (Australian Government). 2020. Australia's AI Ethics Principles. Retrieved August 17, 2022 from https://industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles

[28] Ren Bin Lee Dixon. 2023. A principled governance for emerging AI regimes: lessons from China, the European Union, and the United States. *AI and Ethics* 3 (2023), 793–810.

[29] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. In *Proceedings of the Conference on Robot Learning*. 1–16.

[30] Christof Ebert and Michael Weyrich. 2019. Validation of autonomous systems. *IEEE Software* 36, 5 (2019), 15–23.

[31] Ray Eitel-Porter. 2021. Beyond the promise: Implementing ethical AI. *AI and Ethics* 1, 1 (2021), 73–80.

[32] Iker Esnaola-Gonzalez. 2021. An ontology-based approach for making machine learning systems accountable. *Semantic Web* 1 (2021), 1–5.

[33] Gregory Falco, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, Alwyn Goodloe, Jerry Gupta, Christopher Hart, Marina Jirotka, Henric Johnson, Cara LaPointe, Ashley J. Llorens, Alan K. Mackworth, Carsten Maple, Sigurour Emil Palsson, Frank Pasquale, Alan Winfield, and Zee Kin Yeong. 2021. Governing AI safety through independent audits. *Nature Machine Intelligence* 3, 7 (2021), 566–571.

[34] Gregory Falco and Joshua E. Siegel. 2020. A distributed 'black box' audit trail design specification for connected and automated vehicle data and software assurance. *arXiv preprint arXiv:2002.02780* (2020).

[35] Ming Fan, Le Yu, Sen Chen, Hao Zhou, Xiapu Luo, Shuyue Li, Yang Liu, Jun Liu, and Ting Liu. 2020. An empirical evaluation of GDPR compliance violations in Android mHealth apps. In *Proceedings of the 2020 IEEE 31st International Symposium on Software Reliability Engineering (ISSRE'20)*. IEEE, Los Alamitos, CA, 253–264.

[36] Benjamin Fish and Luke Stark. 2021. Reflexive design for fairness and other human values in formal models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 89–99.

[37] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.* HLS White Paper. Berkman Klein Center for Internet & Society.

[38] Philipp Fukas, Jonas Rebstadt, Florian Remark, and Oliver Thomas. 2021. Developing an artificial intelligence maturity model for auditing. In *Proceedings of the European Conference on Information Systems (ECIS'21)*.

[39] Vahid Garousi, Michael Felderer, and Mika V. Mäntylä. 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology* 106 (2019), 101–121. https://www.sciencedirect.com/science/article/pii/S0950584918301939

[40] Lydia Gauerhof, Richard Hawkins, Chiara Picardi, Colin Paterson, Yuki Hagiwara, and Ibrahim Habli. 2020. Assuring the safety of machine learning for pedestrian detection at crossings. In *Proceedings of the International Conference on Computer Safety, Reliability, and Security*. 197–212.

[41] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM* 64, 12 (2021), 86–92.

[42] Noah J. Goodall. 2014. Machine ethics and automated vehicles. In *Road Vehicle Automation*. Springer, 93–102.

[43] Erika Halme, Ville Vakkuri, Joni Kultanen, Marianna Jantunen, Kai-Kristian Kemell, Rebekah Rousi, and Pekka Abrahamsson. 2021. How to write ethical user stories? Impacts of the ECCOLA method. In *Proceedings of the International Conference on Agile Software Development*. 36–52.

[44] Seung-Ho Han and Ho-Jin Choi. 2022. Checklist for validating trustworthy AI. In *Proceedings of the 2022 IEEE International Conference on Big Data and Smart Computing (BigComp'22)*. IEEE, Los Alamitos, CA, 391–394.

[45] Jette Henderson, Shubham Sharma, Alan Gee, Valeri Alexiev, Steve Draper, Carlos Marin, Yessel Hinojosa, Christine Draper, Michael Perng, Luis Aguirre, Michael Li, Sara Rouhani, Shorya Consul, Susan Michalski, Akarsh Prasad, Mayank Chutani, Aditya Kumar, Shahzad Alam, Prajna Kandarpa, Binnu Jesudasan, Colton Lee, Michael Criscolo, Sinead Williamson, Matt Sanchez, and Joydeep Ghosh. 2021. Certifai: A toolkit for building trust in AI systems. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI'21)*. 5249–5251.

[46] Anne Henriksen, Simon Enni, and Anja Bechmann. 2021. Situated accountability: Ethical principles, certification standards, and explanation methods in applied AI. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 574–585.

[47] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. 2017. Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 95–99.

[48] John N. Hooker and Tae Wan N. Kim. 2018. Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 130–136.

[49] Jennifer Horkoff. 2019. Non-functional requirements for machine learning: Challenges and new directions. In *Proceedings of the 2019 IEEE 27th International Requirements Engineering Conference (RE'19)*. IEEE, Los Alamitos, CA, 386–391.

[50] Fatima Hussain, Rasheed Hussain, Brett Noye, and Salah Sharieh. 2020. Enterprise API security and GDPR compliance: Design and implementation perspective. *IT Professional* 22, 5 (2020), 81–89.

[51] Waqar Hussain, Mojtaba Shahin, Rashina Hoda, Jon Whittle, Harsha Perera, Arif Nurwidyantoro, Rifat Ara Shams, and Gillian Oliver. 2022. How can human values be addressed in agile methods? A case study on SAFe. *IEEE Transactions on Software Engineering* 48, 12 (2022), 5158–5175.

[52] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 560–575.

[53] Javier Camacho Ibáñez and Mónica Villas Olmeda. 2021. Operationalising AI ethics: How are companies bridging the gap between practice and principles? An exploratory study. *AI & SOCIETY* 37 (2021), 1663–1687.

[54] Brian R. Jackson, Ye Ye, James M. Crawford, Michael J. Becich, Somak Roy, Jeffrey R. Botkin, Monica E. de Baca, and Liron Pantanowitz. 2021. The ethics of artificial intelligence in pathology and laboratory medicine: Principles and practice. *Academic Pathology* 8 (2021), 2374289521990784.

[55] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 624–635.

[56] Sophie F. Jentzsch, Sviatlana Höhn, and Nico Hochgeschwender. 2019. Conversational interfaces for explainable AI: A human-centred approach. In *Proceedings of the International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. 77–92.

[57] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[58] Meenu Mary John, Helena Holmström Olsson, and Jan Bosch. 2020. Architecting AI deployment: A systematic review of state-of-the-art and state-of-practice literature. In *Proceedings of the International Conference on Software Business*. 14–29.

[59] Daniel Kasenberg, Thomas Arnold, and Matthias Scheutz. 2018. Norms, rewards, and the intentional stance: Comparing machine learning approaches to ethical training. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 184–190.

[60] B. A. Kitchenham and S. Charters. 2007. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. Technical Report. EBSE.

[61] John C. Knight. 2002. *N*-version programming. In *Encyclopedia of Software Engineering*. Wiley.

[62] Retno Larasati, Anna De Liddo, and Enrico Motta. 2021. AI healthcare system interface: Explanation design for non-expert user trust. In *ACMIUI-WS 2021: Joint Proceedings of the ACM IUI 2021 Workshops*. Vol. 2903. CEUR.

[63] Abolfazl Lavaei, Bingzhuo Zhong, Marco Caccamo, and Majid Zamani. 2021. Towards trustworthy AI: Safe-visor architecture for uncertified controllers in stochastic cyber-physical systems. In *Proceedings of the Workshop on Computation-Aware Algorithmic Design for Cyber-Physical Systems*. 7–8.

[64] Michelle Seng Ah Lee and Jatinder Singh. 2021. Risk identification questionnaire for detecting unintended bias in the machine learning development lifecycle. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 704–714.

[65] Grace A. Lewis, Ipek Ozkaya, and Xiwei Xu. 2021. Software architecture challenges for ML systems. In *Proceedings of the 2021 IEEE International Conference on Software Maintenance and Evolution (ICSME'21)*. IEEE, Los Alamitos, CA, 634–638.

[66] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[67] Q. Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. 2021. Question-driven design process for Explainable AI user experiences. *arXiv preprint arXiv:2104.03483* (2021).

[68] Sin Kit Lo, Yue Liu, Qinghua Lu, Chen Wang, Xiwei Xu, Hye-Young Paik, and Liming Zhu. 2021. Blockchain-based trustworthy federated learning architecture. *arXiv preprint arXiv:2108.06912* (2021).

[69] Sin Kit Lo, Qinghua Lu, Hye-Young Paik, and Liming Zhu. 2021. FLRA: A reference architecture for federated learning systems. In *Proceedings of the European Conference on Software Architecture*. 83–98.

[70] Sin Kit Lo, Qinghua Lu, Liming Zhu, Hye-Young Paik, Xiwei Xu, and Chen Wang. 2022. Architectural patterns for the design of federated learning systems. *Journal of Systems and Software* 191 (2022), 111357.

[71] Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, and Zhenchang Xing. 2022. Towards a roadmap on software engineering for responsible AI. In *Proceedings of the 2022 IEEE/ACM 1st International Conference on AI Engineering: Software Engineering for AI (CAIN'22)*. 101–112.

[72] David D. Luxton. 2014. Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial Intelligence in Medicine* 62, 1 (2014), 1–10.

[73] C. Dianne Martin and Toma Taylor Makoundou. 2017. Taking the high road ethics by design in AI. *ACM Inroads* 8, 4 (2017), 35–37.

[74] Robert C. Martin, Dirk Riehle, and Frank Buschmann. 1997. *Pattern Languages of Program Design 3*. Addison Wesley Longman.

[75] Silverio Martínez-Fernández, Xavier Franch, Andreas Jedlitschka, Marc Oriol, and Adam Trendowicz. 2021. Developing and operating artificial intelligence models in trustworthy autonomous systems. In *Proceedings of the International Conference on Research Challenges in Information Science*. 221–229.

[76] Melissa D. McCradden, Shalmali Joshi, James A. Anderson, Mjaye Mazwi, Anna Goldenberg, and Randi Zlotnik Shaul. 2020. Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *Journal of the American Medical Informatics Association* 27, 12 (2020), 2024–2027.

[77] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54, 6 (2021), 1–35.

[78] Microsoft. 2022. HAX Toolkit. Retrieved August 22, 2022 from https://www.microsoft.com/en-us/haxtoolkit/

[79] Beatriz San Miguel, Aisha Naseer, and Hiroya Inakoshi. 2021. Putting accountability of AI systems into practice. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI'21)*. 5276–5278.

[80] Eric Mitchell, Peter Henderson, Christopher D. Manning, Dan Jurafsky, and Chelsea Finn. 2002. Self-destructing models: Increasing the costs of harmful dual uses in foundation models. In *Proceedings of the 1st Workshop on Pre-Training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*.

[81] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–229.

[82] Henry Muccini and Karthik Vaidhyanathan. 2021. Software architecture for ML-based systems: What exists and what lies ahead. In *Proceedings of the 2021 IEEE/ACM 1st Workshop on AI Engineering–Software Engineering for AI (WAIN'21)*. IEEE, Los Alamitos, CA, 121–128.

[83] Christian Murphy, Gail E. Kaiser, and Marta Arias. 2007. *An Approach to Software Testing of Machine Learning Applications*. Columbia University.

[84] Maskura Nafreen, Saikath Bhattacharya, and Lance Fiondella. 2020. Architecture-based software reliability incorporating fault tolerant machine learning. In *Proceedings of the 2020 Annual Reliability and Maintainability Symposium (RAMS'20)*. IEEE, Los Alamitos, CA, 1–6.

[85] Iman Naja, Milan Markovic, Peter Edwards, and Caitlin Cottrill. 2021. A semantic framework to support AI system accountability and audit. In *Proceedings of the European Semantic Web Conference*. 160–176.

[86] NTIA. 2021. The Minimum Elements for a Software Bill of Materials (SBOM). Retrieved August 18, 2022 from https://www.ntia.doc.gov/files/ntia/publications/sbom_minimum_elements_report.pdf

[87] OECD. 2021. Tools for Trustworthy AI. Retrieved October 14, 2023 from https://www.oecd-ilibrary.org/content/paper/008232ec-en

[88] The United States Department of Commerce. 2021. The Minimum Elements for a Software Bill of Materials (SBOM). Retrieved August 17, 2022 from https://www.ntia.doc.gov/files/ntia/publications/sbom_minimum_elements_report.pdf

[89] Google PAIR. 2021. People + AI Guidebook. Retrieved August 17, 2022 from https://pair.withgoogle.com/guidebook

[90] Harshvardhan Jitendra Pandit, Declan O'Sullivan, and Dave Lewis. 2018. Towards knowledge-based systems for GDPR compliance. In *Proceedings of CKGSemStats@ISWC*.

[91] Emmanouil Papagiannidis, Ida Merete Enholm, Chirstian Dremel, Patrick Mikalef, and John Krogstie. 2021. Deploying AI governance practices: A revelatory case study. In *Proceedings of the Conference on e-Business, e-Services, and e-Society*. 208–219.

[92] Mark C. Paulk, Bill Curtis, Mary Beth Chrissis, and Charles V. Weber. 1993. Capability maturity model, version 1.1. *IEEE Software* 10, 4 (1993), 18–27.

[93] Harsha Perera, Rashina Hoda, Rifat Ara Shams, Arif Nurwidyantoro, Mojtaba Shahin, Waqar Hussain, and Jon Whittle. 2021. The impact of considering human values during requirements engineering activities. *arXiv preprint arXiv:2111.15293* (2021).

[94] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. 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*. 33–44.

[95] Felix Redmill. 2002. Risk analysis—A subjective process. *Engineering Management Journal* 12, 2 (2002), 91–96.

[96] Christoph Regli and Björn Annighoefer. 2022. An anthropomorphic approach to establish an additional layer of trustworthiness of an AI pilot. In *Proceedings of the Software Engineering 2022 Workshops*.

[97] Martin Sand, Juan Manuel Durán, and Karin Rolanda Jongsma. 2022. Responsibility beyond design: Physicians' requirements for ethical medical AI. *Bioethics* 36, 2 (2022), 162–169.

[98] Jana Schaich Borg. 2021. Four investment areas for ethical AI: Transdisciplinary opportunities to close the publication-to-practice gap. *Big Data & Society* 8, 2 (2021), 20539517211040197.

[99] Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. 2020. Principles to practices for responsible AI: Closing the gap. *arXiv preprint arXiv:2006.04707* (2020).

[100] Mario D. Schultz and Peter Seele. 2023. Towards AI ethics' institutionalization: knowledge bridges from business ethics to advance organizational AI ethics. *AI and Ethics* 3 (2023), 99–111.

[101] Kaira Sekiguchi and Koichi Hori. 2020. Organic and dynamic tool for use with knowledge base of AI ethics for promoting engineers' practice of ethical AI design. *AI & SOCIETY* 35, 1 (2020), 51–71.

[102] Mike Shafto, Mike Conroy, Rich Doyle, Ed Glaessgen, Chris Kemp, Jacqueline LeMoigne, and Lui Wang. 2012. Modeling, simulation, information technology & processing roadmap. *National Aeronautics and Space Administration* 32, 2012 (2012), 1–38.

[103] Toby Shevlane. 2022. Structured access to AI capabilities: An emerging paradigm for safe AI deployment. *arXiv preprint arXiv:2201.05159* (2022).

[104] Ben Shneiderman. 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 (2020), Article 26, 31 pages.

[105] Ben Shneiderman. 2021. Responsible AI: Bridging from ethics to practice. *Communications of the ACM* 64, 8 (2021), 32–35.

[106] Vasu Singh, Siva Kumar Sastry Hari, Timothy Tsai, and Mandar Pitale. 2021. Simulation driven design and test for safety of AI based autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 122–128.

[107] Nicolas Six, Andrea Perrichon-Chrétien, and Nicolas Herbaut. 2021. SAIaaS: A blockchain-based solution for secure artificial intelligence as-a-service. In *Proceedings of the International Conference on Deep Learning, Big Data, and Blockchain*. 67–74.

[108] Mona Sloane and Janina Zakrzewski. 2022. German AI start-ups and "AI ethics": Using a social practice lens for assessing and implementing socio-technical innovation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 935–947.

[109] Koen Smit, Martijn Zoet, and John van Meerten. 2020. A review of AI principles in practice. In *Proceedings of the 24th Pacific Asia Conference on Information Systems (PACIS'20)*.

[110] IEEE Computer Society, P. Bourque, and R. Fairley. 2014. *Guide to the Software Engineering Body of Knowledge*. SWEBOK.

[111] Mark Staples, Liming Zhu, and John Grundy. 2016. Continuous validation for data analytics systems. In *Proceedings of the 38th International Conference on Software Engineering Companion*. 769–772.

[112] Nehemia Sugianto, Dian Tjondronegoro, Rosemary Stockdale, and Elizabeth Irenne Yuwono. 2021. Privacy-preserving AI-enabled video surveillance for social distancing: Responsible design and deployment for public spaces. *Information Technology & People*. E-pub ahead of print.

[113] Ahmet Ali Süzen and Mehmet Ali Şimşek. 2020. A novel approach to machine learning application to protection privacy data in healthcare: Federated learning. *Namık Kemal Tıp Dergisi* 8, 1 (2020), 22–30.

[114] Mizuki Takeda, Yasuhisa Hirata, Yueh-Hsuan Weng, Takahiro Katayama, Yasuhide Mizuta, and Atsushi Koujina. 2019. Accountable system design architecture for embodied AI: A focus on physical human support robots. *Advanced Robotics* 33, 23 (2019), 1248–1263.

[115] Steven Umbrello. 2022. The role of engineers in harmonising human values for AI systems design. *Journal of Responsible Technology* 10 (2022), 100031.

[116] Polyxeni Vassilakopoulou. 2020. Sociotechnical approach for accountability by design in AI systems. In *Proceedings of the European Conference on Information Systems (ECIS'20)*.

[117] Wilfried Verachtert, Thomas J. Ashby, Imen Chakroun, Roel Wuyts, Sayantan Das, Sandip Halder, and Philippe Leray. 2021. Privacy preserving amalgamated machine learning for process control. In *Metrology, Inspection, and Process Control for Semiconductor Manufacturing XXXV*. Vol. 11611. SPIE, 329–341.

[118] Andreas Vogelsang and Markus Borg. 2019. Requirements engineering for machine learning: Perspectives from data scientists. In *Proceedings of the 2019 IEEE 27th International Requirements Engineering Conference Workshops (REW'19)*. IEEE, Los Alamitos, CA, 245–251.

[119] Abhishek Wadhwani and Priyank Jain. 2020. Machine learning model cards transparency review: Using model card toolkit. In *Proceedings of the 2020 IEEE Pune Section International Conference (PuneCon'20)*. IEEE, Los Alamitos, CA, 133–137.

[120] Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N. Ahmad Aziz, et al. 2021. Swarm learning for decentralized and confidential clinical machine learning. *Nature* 594, 7862 (2021), 265–270.

[121] Jiasi Weng, Jian Weng, Jilian Zhang, Ming Li, Yue Zhang, and Weiqi Luo. 2019. DeepChain: Auditable and privacy-preserving deep learning with blockchain-based incentive. *IEEE Transactions on Dependable and Secure Computing* 18, 5 (2019), 2438–2455.

[122] Alan F. T. Winfield and Marina Jirotka. 2017. The case for an ethical black box. In *Towards Autonomous Robotic Systems*. Lecture Notes in Computer Science, Vol. 10454. Springer, 262–273.

[123] Wentao Xie and Peng Wu. 2020. Fairness testing of machine learning models using deep reinforcement learning. In *Proceedings of the 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, Los Alamitos, CA, 121–128.

[124] Roland H. C. Yap. 2020. Towards certifying trustworthy machine learning systems. In *Proceedings of the International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning*. 77–82.

[125] Weishan Zhang, Qinghua Lu, Qiuyu Yu, Zhaotong Li, Yue Liu, Sin Kit Lo, Shiping Chen, Xiwei Xu, and Liming Zhu. 2020. Blockchain-based federated learning for device failure detection in industrial IoT. *IEEE Internet of Things Journal* 8, 7 (2020), 5926–5937.

[126] Andi Zhobe, Hamid Jahankhani, Rose Fong, Paul Elevique, and Hassan Baajour. 2021. The magic quadrant: Assessing ethical maturity for artificial intelligence. In *Cybersecurity, Privacy and Freedom Protection in the Connected World*. Springer, 313–326.

[127] Zhibin Zhou, Zhuoshu Li, Yuyang Zhang, and Lingyun Sun. 2022. Transparent-AI blueprint: Developing a conceptual tool to support the design of transparent AI agents. *International Journal of Human–Computer Interaction* 38, 18-20 (2022), 1846–1873.

[128] Liming Zhu, Xiwei Xu, Qinghua Lu, Guido Governatori, and Jon Whittle. 2022. AI and ethics–Operationalizing responsible AI. In *Humanity Driven AI*. Springer, 15–33.

[129] Roberto V. Zicari, Sheraz Ahmed, Julia Amann, Stephan Alexander Braun, John Brodersen, Frédérick Bruneault, James Brusseau, Erik Campano, Megan Coffee, Andreas Dengel, et al. 2021. Co-design of a trustworthy AI system in healthcare: Deep learning based skin lesion classifier. *Frontiers in Human Dynamics* 3 (2021), 688152.

[130] Roberto V. Zicari, John Brodersen, James Brusseau, Boris Düdder, Timo Eichhorn, Todor Ivanov, Georgios Kararigas, Pedro Kringen, Melissa McCullough, Florian Möslein, Naveed Mushtaq, Gemma Roig, Normal Sturtz, Karsten Tolle, Jesmin Jahan Tithi, Irmhild van Halem, and Magnus Westerlund. 2021. Z-Inspection®: A process to assess trustworthy AI. *IEEE Transactions on Technology and Society* 2, 2 (2021), 83–97.

[131] Roberto V. Zicari, James Brusseau, Stig Nikolaj Blomberg, Helle Collatz Christensen, Megan Coffee, Marianna B. Ganapini, Sara Gerke, Thomas Krendl Gilbert, Eleanore Hickman, Elisabeth Hildt, et al. 2021. On assessing trustworthy AI in healthcare. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Frontiers in Human Dynamics* 3 (2021), 30.

Received 13 September 2022; revised 5 June 2023; accepted 26 September 2023