Software engineering for Responsible AI: An empirical study and operationalised patterns

Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, David Douglas, Conrad Sanderson

CSIRO, Australia

***Abstract*—AI ethics principles and guidelines are typically high-level and do not provide concrete guidance on how to develop responsible AI systems. To address this shortcoming, we perform an empirical study involving interviews with 21 scientists and engineers to understand the practitioners' views on AI ethics principles and their implementation. Our major findings are: (1) the current practice is often a done-once-and-forget type of ethical risk assessment at a particular development step, which is not sufficient for highly uncertain and continual learning AI systems; (2) ethical requirements are either omitted or mostly stated as high-level objectives, and not specified explicitly in verifiable way as system outputs or outcomes; (3) although ethical requirements have the characteristics of cross-cutting quality and non-functional requirements amenable to architecture and design analysis, system-level architecture and design are under-explored; (4) there is a strong desire for continuously monitoring and validating AI systems post deployment for ethical requirements but current operation practices provide limited guidance. To address these findings, we suggest a preliminary list of patterns to provide operationalised guidance for developing responsible AI systems.**

***Index Terms*—artificial intelligence, AI, machine learning, responsible AI, ethics, software engineering, software architecture, DevOps**

I. Introduction

Although AI is solving real-world challenges and transforming industries, there are serious concerns about its ability to behave and make decisions in a responsible way. To achieve responsible AI, many AI ethics principles and guidelines have been recently issued by governments, organisations, and enterprises [1], [3], [4]. However, these principles are typically high-level and do not provide concrete guidance on how to develop responsible AI systems.

We first undertake an empirical study involving interviews with 21 AI/ML scientists and engineers at Australia's national scientific research agency (CSIRO), with various backgrounds (such as computer science, health & biosecurity, land & water) and a large variation in the interviewees' degree of experience and responsibility (postgraduate students, research scientists, engineers, team/group leaders) [5]. We asked participants what ethical issues they have considered in their AI/ML projects and how these issues are (or can be) addressed. Australia's AI ethics principles [1] were used as a close-enough representation of the many similar principles around the world [3], [4]. Based on the insights gained from the interviews, literature review, as well as existing software development and design practices, we suggest a preliminary list of process and design patterns applicable to the entire lifecycle of AI systems.

* **Published in:** IEEE/ACM International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), 2022.
DOI: 10.1109/ICSE-SEIP55303.2022.9793864

II. Findings

Table I shows the incidence of themes related to AI ethics principles across the interviews. The top three principles covered in the interviews are *Reliability & Safety*, *Transparency & Explainability*, and *Privacy Protection & Security*. Principles which were covered in roughly half the interviews are *Accountability*, *Human, Societal, Environmental Wellbeing*. The *Human-Centred Values* principle was covered the least in the interviews. Below we summarise the major findings for each of the categories that were identified using open card sorting on interview contents.

- The current practice is a **done-once-and-forget type of risk assessment**, which is likely not to be sufficient for highly uncertain and continual learning AI systems.
- The **inherent trustworthiness** of an AI system for various ethics principles and the **perceived trust** of the system are often mixed in practice. Even for a highly trustworthy AI system, gaining the trust from humans is a challenge that must be addressed carefully for the AI system to be widely accepted.
- **Process and product assurance mechanisms** can be leveraged to achieve trustworthiness for various ethics principles, whereas process and product evidence need to be offered to drive trust.
- Human trust in AI can be improved by attaching **ethics credentials** to AI components/products since the vendors often supply products by assembling commercial and/or open-source AI and non-AI components.
- **An AI model needs to be integrated with the system** to perform the required functions. Combining AI and non-AI components can create new emergent behaviour and dynamics, which require system-level ethical consideration.
- Responsible AI requirements are either **omitted** or **mostly stated as high-level objectives**, and not specified explicitly in a verifiable way as expected system outputs (to be verified/validated) and outcomes (e.g. benefits). Requirements engineering methods need to be extended with ethical aspects for AI systems.

- AI is often complex and hard to explain, thus making detailed risk assessment difficult. **Adopting AI** can be considered as a major architectural design decision when designing a system. Also, an AI component can be designed to be **flexibly switched off** at run-time or changed from decision mode to suggestion mode.
- There are **trade-offs** between many AI ethics principles. The current practice of dealing with the trade-offs is usually the developers following one principle while disregarding the other, rather than building balanced trade-offs with stakeholders making the ultimate value and risk call.
- Although responsible AI requirements have the characteristics of cross-cutting quality and non-functional requirements amenable to **architecture/design analysis and reusable patterns**, they were under-explored in the projects.
- Human-centred approaches have been adopted for **explainability** and **interpretability** taking into account users' background and preferences to improve human trust in AI.
- There is a strong desire for **continuously monitoring and validating AI systems** post deployment for responsible AI requirements, but current MLOps practices provide limited guidance. There is lack of end-to-end development tools to support continuous assurance of AI ethics.
- AI systems usually involve co-evolution of data, model, code, and configurations. **Data / model / code / configuration co-versioning** with model dependency specification is needed to ensure data provenance and traceability.

III. Preliminary List of Patterns

We summarise a preliminary list of operationalised responsible AI assurance **process/design patterns** [2], based on the interview results, literature review, and best practices.

- **Extensible adaptive, dynamic risk assessment**: Ethical risk assessment needs to be automatically performed and adapted for various contexts with certain extension points.
- **Ethics credentials**: Ethics credentials are used to verify the ethical qualification of an organisation, a developer, an AI system, or a component.
- **Bill of materials for AI supply chains**: Bill of materials tracks the supply chain details of the components.
- **Standardised documents**: It is necessary to prepare documents that are compliant with standards and accessible by stakeholders.
- **Integration of requirement-driven and outcome-driven development**: Requirement-driven development and outcome-driven development should be seamlessly integrated for AI systems that are continuously learning based on new data.
- **Verifiable requirements**: Ethical requirements should be specified in a verifiable and measurable way to avoid vague requirements and facilitate requirement validation.
- **AI mode switcher**: The AI component can be activated or deactivated by an AI mode switcher.

TABLE I
Incidence of themes related to AI ethics principles.

Principle	Incidence
Privacy Protection & Security	17 / 21 (81%)
Reliability & Safety	19 / 21 (90%)
Transparency & Explainability	18 / 21 (86%)
Accountability	13 / 21 (62%)
Contestability	8 / 21 (38%)
Fairness	10 / 21 (47%)
Human-Centred Values	3 / 21 (14%)
Human/Societal/Environmental Wellbeing	11 / 21 (52%)

- **Decision mode switcher**: Decision mode switcher provisions various decision-making modes, i.e. fully automatic or suggestion with human-in-the-loop (kill switch, override, fallback).
- **Federated learner**: Federated learning can be viewed as an architectural pattern that performs model training locally at each client and ensembles the model update centrally.
- **Co-architecting of AI and non-AI components**: AI systems require co-architecting of both AI and non-AI components to meet system-level as well as model-level ethical requirements.
- **System-level simulation**: Simulation improves the understanding of system behaviour and reduces ethical risk.
- **Human-centred explainable interface**: Explainability can be improved by a human-centered interface that explains models and interprets results to various stakeholders.
- **Construction for/with reuse**: Both AI model pipeline code and system component code can be reused to improve reliability.
- **Ethical acceptance tests**: Ethical requirements can be verified through ethical acceptance tests.
- **Model deployer**: Model deployer offers various deployment strategies, such as multiple models and online learning.
- **Audit black box**: Ethical metric data can be captured at run time by a black box for auditing purposes.
- **Continuous validator**: Ethical metrics need to be identified and continuously validated at run-time.

References

[1] Australian Government Department of Industry, Science, Energy and Resources. Australia's AI Ethics Principles, 2020. URL: https://industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles. Accessed: 04 Oct 2021.

[2] L. Bass, P. Clements, and R. Kazman. *Software Architecture in Practice*. Addison-Wesley Professional, 4th edition, 2021.

[3] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*, (2020-1), 2020.

[4] A. Jobin, M. Ienca, and E. Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

[5] C. Sanderson, D. Douglas, Q. Lu, E. Schleiger, J. Whittle, J. Lacey, G. Newnham, S. Hajkowicz, C. Robinson, and D. Hansen. AI ethics principles in practice: Perspectives of designers and developers. arXiv: 2112.07467, 2021.