Although artificial intelligence (AI) has significant potential and capacity to stimulate economic growth and improve productivity across a growing range of domains, there are serious concerns about AI systems’ ability to behave and make decisions in a responsible manner. According to Gartner’s recent report,\(^1\) 21% of organizations have already deployed or plan to deploy responsible AI technologies within the next 12 months.

**Introduction**

Many ethical principles and guidelines have been recently issued by governments, research institutions, and companies.\(^2\) However, these principles are high level and can hardly be used in practice by developers. Responsible AI research has been focusing on algorithm solutions limited to a subset of issues, such as fairness.\(^3\) Ethical issues can enter at any point of the software engineering lifecycle and are often at the system level, crosscutting many components of AI systems. To try to fill the principle–algorithm gap, some development guidelines have started to appear. However, those efforts tend to be high-level development process checklists\(^4,5\) and ad hoc sets lacking state-related linkages for final products.\(^6\)

Therefore, in this article, rather than staying at the ethical principle level or AI algorithm level, we take a pattern-oriented approach and focus on system-level design patterns to build responsible-AI-by-design into final AI products. The design patterns are collected based on the results of a systematic literature review (SLR) and can be embedded into the design of AI systems as product features to contribute to responsible-AI-by-design. We identify the lifecycle of a provisioned AI system in which the states or state transitions are associated with design patterns to show when the design patterns can take effect. The lifecycle along with the design pattern annotations provide a responsible-AI-focused view of system interactions and a guide to effect the use of design patterns to implement responsible AI from a

---

**Responsible-AI-by-Design**

A Pattern Collection for Designing Responsible Artificial Intelligence Systems

Qinghua Lu, Liming Zhu, Xiwei Xu, and Jon Whittle, Data61, CSIRO

---

**Digital Object Identifier** 10.1109/MS.2022.3233582

**Date of current version:** 18 April 2023
system perspective. To the best of our knowledge, this is the first study that provides concrete and actionable system-level design guidance for architects and developers to reference.

**Methodology**

To operationalize responsible AI, we performed an SLR to identify design patterns that architects and developers can use during the development process. Figure 1 illustrates the methodology. The research question is as follows: “What solutions for responsible AI can be identified?” The research question focuses on identifying the reusable patterns for responsible AI. We used “AI,” “responsible,” and “solution” as the key terms and included synonyms and abbreviations as supplementary terms to increase the search results. The main data sources are the Association for Computing Machinery Digital Library, IEEE Xplore, Science Direct, Springer Link, and Google Scholar. The study only includes papers and articles that present concrete design or process solutions for responsible AI and excludes papers and articles that only discuss high-level frameworks. A set of 159 primary studies was identified. The complete SLR protocol is available at https://drive.google.com/file/d/1Ty4Cpj_GzePzwxw5jGK5ZS5AvK3Ay3Q/view?usp=sharing. We use the ethical principles listed in Harvard University’s mapping study: privacy, accountability (professional responsibility is merged into accountability due to the overlapping definitions), safety and security, transparency and explainability, fairness and nondiscrimination, human control of technology, and promotion of human values.

**Lifecycle of a Provisioned AI System**

Figure 2 illustrates the lifecycle of a provisioned AI system using a state diagram and highlights the patterns associated with relevant states or transitions, which show when the design patterns could take effect. We have limited the scope to the design patterns that can be embedded into AI systems and the provisioned supply chain tool chain as final product features. The best practices of the development process, including some patterns related to offline model training, are out of the scope of this article.

Before an AI system is provisioned, the supply chain information can be accessed through the bill of materials. Users can be required to provide the verifiable ethical credentials to show their capability to operate the systems, and users can examine the
system’s verifiable ethical credentials for ethical compliance checking. Once the AI system starts serving, it is important to perform system-level simulation through an ethical digital twin. An ethical sandbox can be used to physically separate AI components from non-AI components.

When an AI system is requested to execute a task, decision making is often needed before executing the task. An AI component can be activated or deactivated through an AI model switcher to automatically make the decision or involve human experts to review the suggestion. A multimodel decision maker can use different models to make a single decision and cross-check the results. Similarly, homogeneous redundancy can be applied to the system design to enable fault tolerance.

Both the behaviors and decision-making outcomes of the AI system are monitored and validated through a continuous ethical validator. Incentives for ethical behaviors can be maintained by an incentive registry. If the system fails to meet the requirements (including the ethical requirements) or a near miss is detected, the system needs to be updated. A federated learner retrains the model locally at each client to protect data privacy. The co-versioning registry can be used to track the coevolution of AI system components or assets.

An ethical knowledge base can be built to make the ethical knowledge systematically accessed and used when developing or updating the AI system. The AI system needs to be audited regularly or when major failures/near misses occur. An ethical black box can be designed to record the critical data that can be kept as evidence. A global-view auditor can be built on top to provide global-view accountability when multiple systems are involved in an accident. The stakeholders can determine whether to abandon the AI system if it no longer fulfills the requirements.

**Design Patterns**

To operationalize responsible AI, Figure 3 lists a collection of patterns for responsible-AI-by-design. The full version of design patterns is available at https://drive.google.com/file/d/1SBuqkdx91hzczxiGjtxxMyl5JzIVBK6/view?usp=sharing.

**FIGURE 2.** The lifecycle of a provisioned AI system.
### FIGURE 3. Operationalized design patterns for responsible AI systems.

| Pattern Name | Problem | Context | Consequences | Related patterns |
|--------------|---------|---------|--------------|------------------|
| Ethical sandbox | Trustworthiness | Asynchronous/hierarchical/decentralised | - | - |
| Ethical black box | Trustworthiness | Homogeneous ethical knowledge base | - | - |
| Immune log | Trustworthiness | A federation of responder registries and global-view auditors | - | - |
| Ethical twin | Trustworthiness | Multimodel skills, increased training, decreased development effort | - | - |

**Patterns Table**

| Pattern Name | Problem | Context | Consequences | Related patterns |
|--------------|---------|---------|--------------|------------------|
| Ethical sandbox | Trustworthiness | Asynchronous/hierarchical/decentralised | - | - |
| Ethical black box | Trustworthiness | Homogeneous ethical knowledge base | - | - |
| Immune log | Trustworthiness | A federation of responder registries and global-view auditors | - | - |
| Ethical twin | Trustworthiness | Multimodel skills, increased training, decreased development effort | - | - |

**Solution**

- Publicly accessible design documents and codebase
- Bill of materials for all materials and related documentation
- Immutable data infrastructure
- Federated learner, immutable data infrastructure
- Ethical digital twin, continuous ethical knowledge base

**Benefits**

- Increased trust, AI human-in-the-loop, increased management effort
- Increased security, increased trust, AI human-in-the-loop, increased management effort
- Increased trust, AI human-in-the-loop, increased management effort
- Increased trust, AI human-in-the-loop, increased management effort

**Drawbacks**

- Increased development effort, required more multimodel skills, decreased training effort
- Increased operating cost, increased ethical quality, decreased performance penalty
- Increased ethical quality, decreased performance penalty
- Increased ethical quality, decreased performance penalty
- Increased ethical quality, decreased performance penalty

**REFERENCES**

1. Charles, S. (2021). Accountability and explainability in AI: A systematic literature review. IEEE Software, 38(2), 6-13.
2. von der Gracht, F., & Smith, M. (2020). The ethics of artificial intelligence: A framework for responsible AI. IEEE Software, 37(3), 6-13.
Bill of Materials
AI product vendors often create AI systems by assembling commercial or open source AI and/or non-AI components from third parties. AI users often have ethical concerns about the procured AI systems/components. Before an AI system is provisioned, the supply chain information can be accessed through the bill of materials, which keeps a formal machine-readable record of the supply chain details of the components used in building an AI system, such as the component name, version, supplier, dependency relationship, author, and timestamp. The purpose of the bill of materials is to provide traceability and transparency into the components that make up AI systems so that ethical issues can be tracked and addressed. There have been many tools to generate a software bill of materials for practitioners, such as Dependency-Track. To ensure traceability and integrity, immutable data infrastructure is needed to store the data of the bill of materials. For example, the manufacturers of autonomous vehicles can maintain a material registry contract on blockchain to track their components’ supply chain information, e.g., the version and supplier of the third-party AI-based navigation component.

Verifiable Ethical Credentials
Verifiable ethical credentials are cryptographically verifiable data that can be used as strong proof of ethical compliance for AI systems, components, artifacts, and stakeholders (such as developers and users). Before using the provisioned AI systems, users verify the systems’ ethical credentials to check if the systems are compliant with AI ethics principles or regulations. On the other hand, users are often required to provide the ethical credentials to use and operate the AI systems. Publicly accessible data infrastructure needs to be built to support the generation and verification of ethical credentials. For example, before driving a vehicle, the driver is requested to scan her/his ethical credential to show she/he has the capability to drive safely while verifying the ethical credential of the vehicle’s automated driving system shown on the center console.

Ethical Digital Twin
Before running the provisioned AI system in a production environment, it is critical to conduct system-level simulation through an ethical digital twin running on a simulation platform to monitor the behaviors of AI systems and predict potential ethical risks. An ethical digital twin can also be designed as a component at the operation infrastructure level to examine an AI system’s runtime behaviors and decisions based on the abstract simulation model using real-world data. The risk assessment results can be used by the system or users to take further actions to mitigate the potential ethical risk. For example, the manufacturers of autonomous vehicles can use the ethical digital twin to explore the limits of autonomous vehicles based on the collected runtime data, such as NVIDIA DRIVE Sim and xFpro.

Ethical Sandbox
It is risky to execute the whole system including AI components and non-AI components in the same environment. When an AI system is being served, an ethical sandbox can be used to physically separate AI components from non-AI components by running the AI components in a self-contained emulation execution environment, e.g., sandboxing the unverified visual perception component. The AI components placed in the ethical sandbox have no access to the rest of the AI system. All of the hardware and software functionality of the AI components are duplicated in the ethical sandbox.

Thus, the AI components can run safely under supervision before being deployed at scale. For example, Fastcase AI Sandbox provides a secure AI execution platform for analyzing data safely in a secure environment. The maximal tolerable probability should be set as an ethical margin for the sandbox against the ethical requirements. A watch dog can be added to restrict the execution time of the AI component to avoid the potential ethical risk, e.g., only executing the visual perception component for 10 min on roads designed especially for autonomous vehicles.

AI Mode Switcher
When to activate AI is a major architectural design decision when designing a software system. When an AI system is making a decision, the AI mode switcher enables efficient invocation and dismissal mechanisms for activating or stopping the AI component when needed. The kill switch is a special type of invocation mechanism that immediately turns off the AI component and terminates its negative effects, e.g., switching off the autopilot functionality and its Internet connection. The AI component can make decisions automatically or provide suggestions to human experts in high-risk situations. The decisions can be approved or overridden by a human expert (e.g., skipping the path suggested by the navigation system). If the system state after acting on an AI decision is not expected by human experts, a fallback can be triggered to reverse
the system back to the previous state. A built-in guard ensures that the AI component is only being used under the predefined risk categories.

Multimodel Decision Maker
The reliability of traditional software is dependent on the design of software components. One of the reliability practices in the reliability community is redundancy, which can be applied to AI components. When decisions are being made by an AI system, a multimodel decision maker can run different models to make a single decision, e.g., using different algorithms for visual perception. Reliability can be improved by using different models under different contexts (e.g., different user groups or regions). In addition, fault tolerance can be enabled by cross-checking the results given by multiple models (e.g., only accepting the same results from the deployed models). IBM Watson Natural Language Understanding makes predictions using an ensemble learning framework that includes multiple emotion-detection models.

Homogeneous Redundancy
Ethical failures in AI systems can cause serious damage to humans or the environment. N-version programming is a design pattern for dealing with the reliability issues of traditional software. This concept can be adapted and applied to AI system design. Homogeneous redundancy (e.g., two brake control components) can be applied to tolerate the highly uncertain AI system components that can make unethical decisions or the adversary hardware components that produce malicious data or behave unethically. When an AI system is executing a task, a cross-check can be performed for the outputs given by multiple redundant components of a single type.

Incentive Registry
Incentives are effective for motivating AI systems to execute tasks in a responsible manner. When executing a task, an incentive registry records the rewards that are given for the decisions and behaviors of AI systems, e.g., rewards for the recommended path without safety risk. There are different ways to enforce the incentive mechanism, e.g., designing the incentive mechanism on blockchain-based data infrastructure that is publicly accessible using reinforcement learning. However, it is challenging to design the mechanisms in a responsible AI context since it is difficult to measure the ethical impact of the decisions and behaviors of AI systems on some ethical principles (such as human values). Furthermore, consensus needs to be reached on the incentive mechanism by all the stakeholders. Additionally, in some cases, ethical principles conflict with each other, making the design of an incentive mechanism harder. FLowBC is a tool that uses blockchain to incentivize training contributions for federated learning.

Continuous Ethical Validator
AI systems often need to conduct continual learning when data drift or unethical behavior is detected in production. When an AI system executes tasks, a continuous ethical validator monitors and validates the outcomes of AI systems (e.g., the path suggested by the navigation system) against the ethical requirements. The outcomes of AI systems are the consequences of the decisions and behaviors of the systems, i.e., whether the AI system behaves ethically or provides the promised benefits in a given situation. The time and frequency of validation can be predefined within the continuous validator. Version-based feedback and a rebuild alert can be sent when the ethical requirements are met or breached. An incentive registry can be used to reward or punish the ethical/unethical behavior or decisions of AI systems.

Ethical Knowledge Base
AI systems involve broad ethical knowledge, including AI ethics principles, regulations, unethical use cases, etc. Unfortunately, such ethical knowledge is scattered in different documents (e.g., AI incidents) and is usually implicit or even unknown to developers, who primarily focus on the technical aspects of AI systems and do not have an ethics background. This results in negligence or the ad hoc use of relevant ethical knowledge in AI system development. An ethical knowledge base is built upon a knowledge graph to make meaningful entities, concepts, and their rich semantic relationships explicit and traceable across heterogeneous documents so that the ethical knowledge can be systematically accessed, analyzed, and used when developing or updating AI systems. For example, an ethical knowledge base can be used to support continuous ethical risk assessment. An ethical knowledge base can be built based on AI ethics principles and frameworks, as well as actual AI use cases discussed in existing papers and articles.

Co-Versioning Registry
AI systems involve different levels of dependencies and need frequent evolution when data drift or unethical behavior occurs. Co-versioning of the components of AI systems or AI assets generated in AI pipelines provides provenance guarantees across the entire lifecycle of AI systems. There have been many version control tools for managing the co-versioning of data and models, such as DVC. When updating an AI system, a co-versioning
registry can track the co-evolution of components or AI assets. There are different levels of co-versioning: co-versioning of AI components and non-AI components as well as co-versioning of the assets within the AI components (i.e., co-versioning of the data, model, code, and configurations). A publicly accessible data infrastructure can be used to maintain the co-versioning registry to provide a trustworthy trace for dependencies. For example, a co-versioning registry contract can be built on blockchain to manage different versions of visual perception models and the corresponding training datasets.

**Federated Learner**

Despite the widely deployed mobile or Internet of Things devices generating massive amounts of data, data hungriness is still a challenge, given the increasing concerns regarding data privacy. When learning or updating AI models, a federated learner preserves the data privacy by performing the model training locally on the client devices and formulating a global model on a central server based on the local model updates, e.g., training the visual perception model locally in each vehicle. Decentralized learning is an alternative to federated learning that uses blockchain to remove the single point of failure and coordinate the learning process in a fully decentralized way. In the event of negative outcomes, the responsible humans can be traced and identified by an ethical black box for accountability.

**Ethical Black Box**

The black box was introduced initially for aircraft several decades ago for recording critical flight data. The purpose of embedding an ethical black box in an AI system is to audit an AI system and investigate why and how the 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. For example, an ethical black box could be built into the automated driving system to record the behaviors of the system and driver and their effects. 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 cloud-based data storage).

**Global-View Auditor**

There can be more than one AI system involved in an ethical incident (e.g., multiple autonomous vehicles in a car accident). During auditing, it is often challenging to identify the liability, as the data collected from each of the involved systems can conflict with each other. A global-view auditor can enable accountability by analyzing the data discrepancies between the involved AI systems and identifying the liability for the ethical incident. This pattern can also be applied to improve the reliability of an AI system by taking the data from other systems. For example, an autonomous vehicle increases its visibility using the perception data collected from other vehicles. All of the historical data of AI systems can be recorded by an immutable log for third-party auditing.

**References**

1. “IT budgets are growing. Here’s where the money’s going.” Gartner. Accessed: Jan. 16, 2023. [Online]. Available: https://www.gartner.com/en/articles/it-budgets-are-growing.-here-s-where-the-money-s-going
2. A. Jobin, M. Ienca, and E. Vayena, “The global landscape of AI ethics guidelines,” Nature Mach. Intell., vol. 1, no. 9, pp. 389–399, Sep. 2019. doi: 10.1038/s42256-019-0088-2
3. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” ACM Comput. Surveys, vol. 54, no. 6, pp. 1–35, Jul. 2022, doi: 10.1145/3457607.
4. “Ethics by design and ethics of use approaches for artificial intelligence,” European Commission, Brussels, Belgium, Nov. 2021. [Online]. Available: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf
5. “Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems,” Creative Commons, Mountain View, CA, USA, 2017. [Online]. Available: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf
6. Q. Lu, L. Zhu, X. Xu, J. Whipple, D. Douglas, and C. Sanderson, “Software engineering for responsible AI: An empirical study and operationalised patterns,” in Proc. IEEE/ACM 44th Int. Conf. Softw. Eng., Softw. Eng. Pract., 2022, pp. 241–242, doi: 10.1145/3510457.3513063.
7. J. Fjeld et al. “Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI.” SSRN. Accessed: 2020. [Online]. Available: https://ssrn.com/abstract=3518482
8. “The minimum elements for a software bill of materials (SBOM),” U.S. Dept. Commerce, Washington, DC, USA, Jul. 2021. [Online]. Available: https://www.ntia.doc.gov/files/ntia/publications/sbom_minimum_elements_report.pdf
9. Continuous SBOM Analysis Platform. OWASP Dependency-Track. Accessed: Jan. 16, 2023. [Online]. Available: https://dependencytrack.org
10. I. Barclay, A. Preece, I. Taylor, S. K. Radha, and J. Nabrzyski, “Providing assurance and scrutability on shared data and machine learning models with verifiable credentials,” Concurreny Comput. Pract. Exp., early access, 2022, doi: 10.1002/cpe.6997.
11. W. Chu, “A decentralized approach towards responsible AI in social ecosystems,” 2021, arXiv:2102.06362.
12. “Building trusted identity networks,” SecureKey, Toronto, ON, Canada. Accessed: Jan. 16, 2023. [Online]. Available: https://securekey.com
13. “NVIDIA DRIVE sim - Powered by omniverse.” Nvidia. Accessed: Jan. 16, 2023. [Online]. Available: https://developer.nvidia.com/drive/drive-sim
14. rFpro Trailer: Welcome to Our World of Driving Simulation. rFpro. Accessed: Jan. 16, 2023. [Online Video]. Available: https://rfpro.com
15. A. Lavaei, B. Zhong, M. Caccamo, and M. Zamani, “Towards trustworthy AI: Safe-visor architecture for uncertified controllers in stochastic cyber-physical systems,” in Proc. Workshop Comput.-Aware Algorithmic...
16. “Fastcase AI sandbox.” Fastcase. Accessed: Jan. 16, 2023. [Online]. Available: https://www.fastcase.com/sandbox/

17. “Future of driving.” Tesla. Accessed: Jan. 16, 2023. [Online]. Available: https://www.tesla.com/autopilot

18. Tutorials Multi-Model Training for Tensorflow. (2018). NeuroAILab. [Online]. Available: http://neuroailab.stanford.edu/tutorials/fundamentals/multimodel.html

19. “Watson natural language understanding: Natural language processing for advanced text analysis.” IBM. Accessed: Jan. 16, 2023. [Online]. Available: https://www.ibm.com/au-en/cloud/watson-natural-language-understanding

20. L. N. Tidjon and F. Khomh, “Threat assessment in machine learning based systems,” 2022, arXiv:2207.00091.

21. J. Weng, J. Weng, J. Zhang, M. Li, Y. Zhang, and W. Luo, “Deepchain: Auditable and privacy-preserving deep learning with blockchain-based incentive,” IEEE Trans. Dependable Secure Comput., vol. 18, no. 5, pp. 2438–2455, Sep./Oct. 2021, doi: 10.1109/TDSC.2019.2952332.

22. “Oschart/FLoBC.” GitHub. Accessed: Jan. 16, 2023. [Online]. Available: https://github.com/Oschart/FLoBC

23. I. Naja, M. Markovic, P. Edwards, and C. Cottrill, “A semantic framework to support AI system accountability and audit,” in Proc. Eur. Semantic Web Conf., 2021, pp. 160–176.

24. Open-Source Version Control System for Machine Learning Projects. Data Version Control. Accessed: Jan. 16, 2023. [Online]. Available: https://dvc.org/

25. S. Caldas et al. (2019). “LEAF: A benchmark for federated settings.” Presented at Workshop on Federated Learning for Data Privacy and Confidentiality. [Online]. Available: https://leaf.cmu.edu

26. G. Falco and J. E. Siegel, “A distributed ‘black box’ audit trail design specification for connected and automated vehicle data and software assurance,” 2020, arXiv:2002.02780.

27. B. S. Miguel, A. Naseer, and H. Inakoshi, “Putting accountability of AI systems into practice,” in Proc. 29th Int. Joint Conf. Artif. Intell., 2021, pp. 5276–5278, doi: 10.24963/ijcai.2020/768.