From Design to Deployment: Accelerating Responsible AI adoption with MLOps and Design Thinking

Written by Cecilia Nunn,
Responsible AI lead at
Digital Catapult
for the Innovate UK
BridgeAI programme

February 2025

UKRI Innovate UK
BridgeAI

Abstract

Implementing responsible Artificial Intelligence (AI) in businesses remains challenging due to the perception that ethical principles are disjointed, abstract, and unintuitive within technical workflows. For AI ethics to gain widespread adoption in industry, these principles must be presented in formats that engineers and data scientists can easily understand and integrate into their routines.

One effective approach is leveraging the increasingly adopted Machine Learning Operations (MLOps) framework to promote responsible AI (RAI) practices. Integrating responsible AI principles into MLOps workflows yields mutual benefits, including enhanced fairness, accessibility, bias reduction, improved risk mitigation, and increased consumer confidence through greater transparency and explainability. To ensure responsible MLOps remains adaptable to evolving AI environments and regulatory requirements, its adoption should be informed by a design-thinking approach to ensure it stays relevant and user-friendly.

Introduction

Creating an AI system that balances accuracy, improved decision-making, cost reduction, enhanced productivity, and scalability is inherently complex. This complexity is heightened by the need to navigate ambiguous ethical standards and the impending regulatory changes introduced by the EU AI Act, which together contribute to industry uncertainty. Building consumer trust and encouraging the widespread adoption of AI products and systems requires embedding ethical principles directly into their design and operation. Long-term success depends on businesses integrating key values such as fairness, accountability, safety, privacy, and transparency into their AI systems. This paper explores the challenges of implementing responsible AI (RAI) in technical and practical contexts and proposes incorporating RAI principles into an MLOps framework.

Using MLOps as a vehicle for RAI provides a solution to operationalising abstract principles into tangible actions, thereby protecting brand reputation and future AI investment. This paper emphasised the importance of adopting a user-centred, design-thinking approach to RAI, which shifts the focus from general ethical AI concepts to addressing specific multifaceted RAI challenges. It argues for the development of tailored tools that align with users' needs and project-specific requirements across the AI lifecycle. By integrating a design-thinking perspective with RAI and MLOps, this paper examines how RAI principles can be effectively operationalised throughout the MLOps lifecycle. It analyses various RAI tools and their application at different stages of AI design, development, and deployment, ensuring the creation of systems that are not only technically robust but also responsible and trustworthy (Aragon et al., 2022).

Background

Responsible AI

Responsible AI encompasses diverse principles, practices, and methodologies aimed at aligning AI development with ethical standards, and regulatory frameworks. With impending regulatory changes, technology organisations must learn to adapt to adhere to these principles. Despite over 250 ethical AI guidelines that have been established by various organisations across different industries and contexts, their core principles—fairness, transparency, explainability, reliability, privacy, and trustworthiness—remain consistent (Billeter et al., 2024). As AI products and services become increasingly prevalent in society, there have been numerous high-profile cases where the lack of responsible AI considerations has led to adverse outcomes. For example, Generative AI (Gen AI) Large Language Models (LLMs) and images generate biased, sexist, and racist responses. Addressing these challenges requires RAI to be recognised as an essential, non-negotiable component of AI development (Zhu et al., 2022). To rebuild trust in technology, countries are establishing guidelines to foster responsible experimentation, reduce risks, and enhance public confidence in AI systems (World Economic Forum, 2023).

As RAI's inclusion becomes more widely accepted, it is paramount that RAI is easily understood, adopted, and relevant for every industry. To date, there have been criticisms about RAI principles and guidelines being abstract, high level and lacking concrete technical application (Billeter et al.,2024). They have also been critiqued as being self-serving to organisations who do not want to commit to change and ultimately allow 'ethics washing' as they become a marketing campaign to create a false sense of security in products (Bietti,2020; Diaz-Rodriguez et al.,2023). As a result, countries are introducing new regulations and exerting greater pressure on technology companies to move beyond these boundaries. However, these upcoming agreements vary in requirements and legal ramifications. Such as the voluntary 'Frontier AI Safety Commitments' (UK Government, 2024), the forthcoming legally binding EU AI Act (European Parliament, 2024) and the UK's Online Safety Act (UK Government, 2023). In recognition of the challenges discussed, RAI must provide mechanisms in which the approach can become ingrained into technical guidelines and workflows.

Design-thinking RAI

Applied technical advancements must be underpinned by a robust technical framework that also supports a holistic approach aligned with RAI principles. Several related concepts have gained traction since the late 2010s, building on earlier research while aligning with RAI goals. These include Explainable AI (XAI), Trustworthy AI (TAI), responsible AI governance, Robust AI, and AI interpretability (AII). Although this paper does not discuss these concepts in detail, they demonstrate the broad spectrum of advancements within the industry. Nevertheless, it is evident there is no one-size-fits-all solution or 'silver bullet' for implementing RAI. While similarities exist among AI products, models, and solutions, each presents unique challenges depending on factors such as stakeholders, users, risks, and available data. For RAI to be impactful, it is essential to consider the perspective from which these challenges are addressed. This paper argues that design thinking provides a valuable framework for guiding the design and development of AI products (see Figure 1). The following section will explore how and why this methodology is effective.

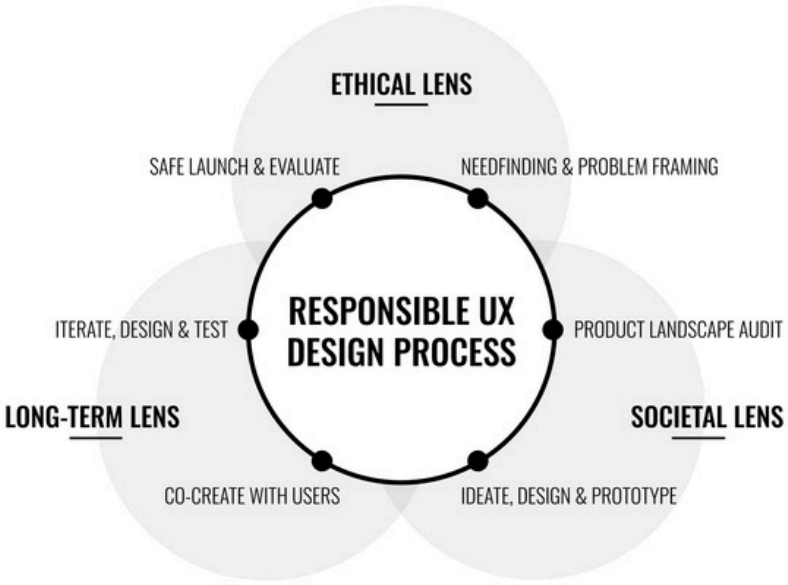


Figure 1 - Source: (Çakmakli, 2024)
This figure shows the design process used for UX design. This report argues using this framework in the context of operationalising RAI.

Applying design thinking to AI development shifts the focus from merely following industry trends to solving real problems. To remain competitive, the question evolves from "How can we build an AI project?", to "What challenges do we face, could AI provide a viable solution, and how can we test this hypothesis?" (Paton and Dorst, 2011). A reactive strategy based on minimal effort and conventional problem-solving is common in the first approach, frequently resulting in solutions that lack clear definition, have limited understanding of users, and are difficult to access or use as they are not solving the right problem. In contrast, the latter approach, emphasises deconstructing the problem and adopting a broader exploratory perspective, ensuring the root issue is addressed and solutions are not rushed (Dorst, 2011).

A significant factor in AI projects failing to reach deployment is not only their inherent complexity but also the tendency to attempt to solve the wrong problems from the outset (Davenport et al., 2018). By the time deployment approaches, organisations may find the model is irrelevant to their needs or that the AI solution does not relate to the original problem (Brown,2009). Approaches designed to meet the growing demand for more technical methods of implementing responsible AI (RAI) must be grounded in a design-thinking framework. This framework should integrate iterative testing, stakeholder involvement, and a user-centred focus throughout development. Without such grounding, RAI approaches risk remaining abstract, vague, and impractical, detached from the realities of AI deployment (Jobin et al., 2019). With this, the solution can be focused, accessible and useful to its users and impactful in its context. Having outlined the user-centred and design-thinking lens that RAI should adopt and will now explore how it can be integrated with existing technical workflows to make its application more practical, usable, and accessible.

Machine Learning Operations (MLOps)

AI researchers and practitioners are increasingly exploring the rapidly evolving field of MLOps. They have recognised its role in enhancing the deployment of ML models and delivering sustainable business value (Godwin & Melvin, 20204). The term MLOps (Machine Learning Operations) doesn't have a formal definition but it builds on the concept of DevOps (Development Operations) to integrate machine learning. It was first coined in 2015 in the research paper 'Hidden Technical Debts in Machine Learning Systems' which highlighted the challenges experienced by teams when developing and deploying ML systems and the need for reflecting on effective planning of ML models (Sculley et al., 2015). MLOps provides tools, practices and methods for collaborative software development tailored to each team's needs. Whilst teams work rapidly on developing new AI tools, MLOps offers an accelerated systematic approach that is vital for a clear, concise assembly line at each ML life-cycle checkpoint (see Figure 2).

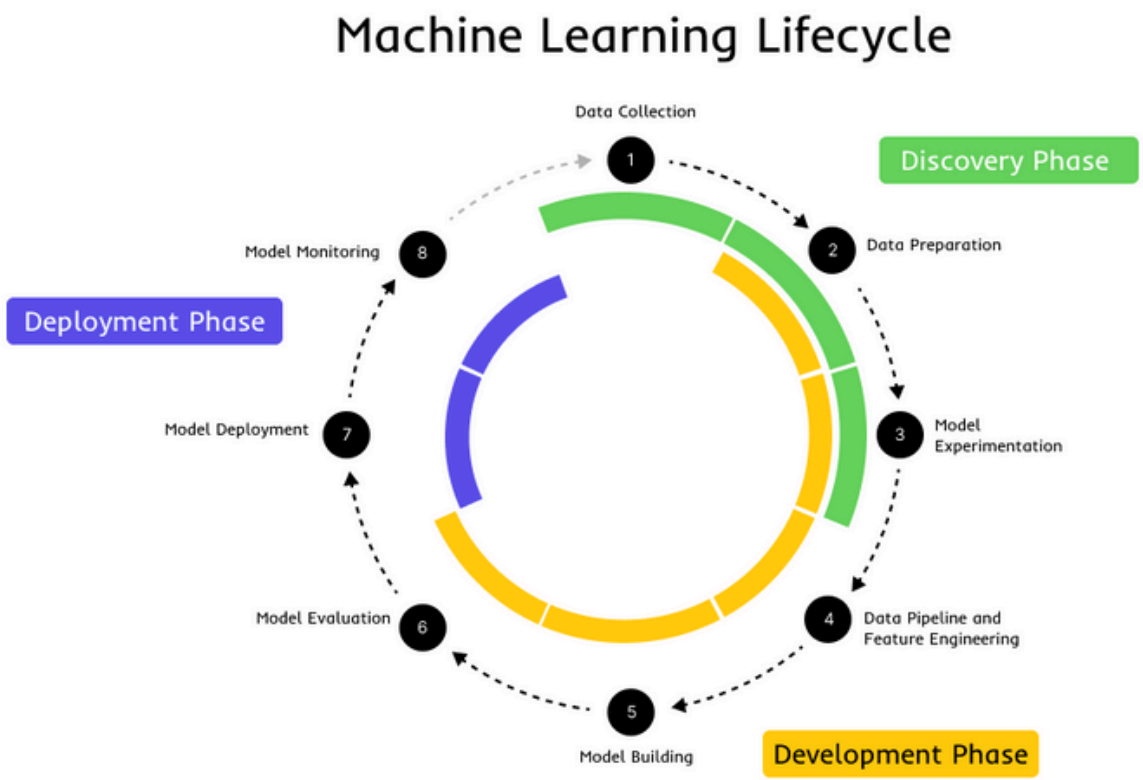


Figure 2 - Source: (Lunardi, 2024)

A visual representation of the Machine Learning Lifecycle, including the key phases and checkpoints which are included in MLOps tools.

According to Rexter Analytics 2023 Data Science survey, only 32% of AI/ML models successfully move from pilot to deployment and production. Among the significant challenges that organisations face are a lack of clear strategy, a lack of cross-functional collaboration, unclear goals and metrics, starting with models that are too large, unaligned and unclear business objectives, and failure to plan for scalability (Treveil et al., 2020). The complexity of ML systems extends beyond the code, which, while significant for tasks such as item classification or value prediction, represents only one aspect. Other complex and nuanced components also play a crucial role in the successful development and navigation of ML systems. Despite significant advancements in MLOps since its inception, much of the discourse remains limited to broad characteristics (see Figure 3) and the tools used, with insufficient attention given to addressing the specific challenges outlined above (Matsui and Goya, 2022). While optimising the ML lifecycle and improving its stability and reliability within the software delivery process would deliver significant benefits to end users, a critical gap persists in understanding how to effectively implement these improvements at each stage of the lifecycle (Tamburri, 2020).

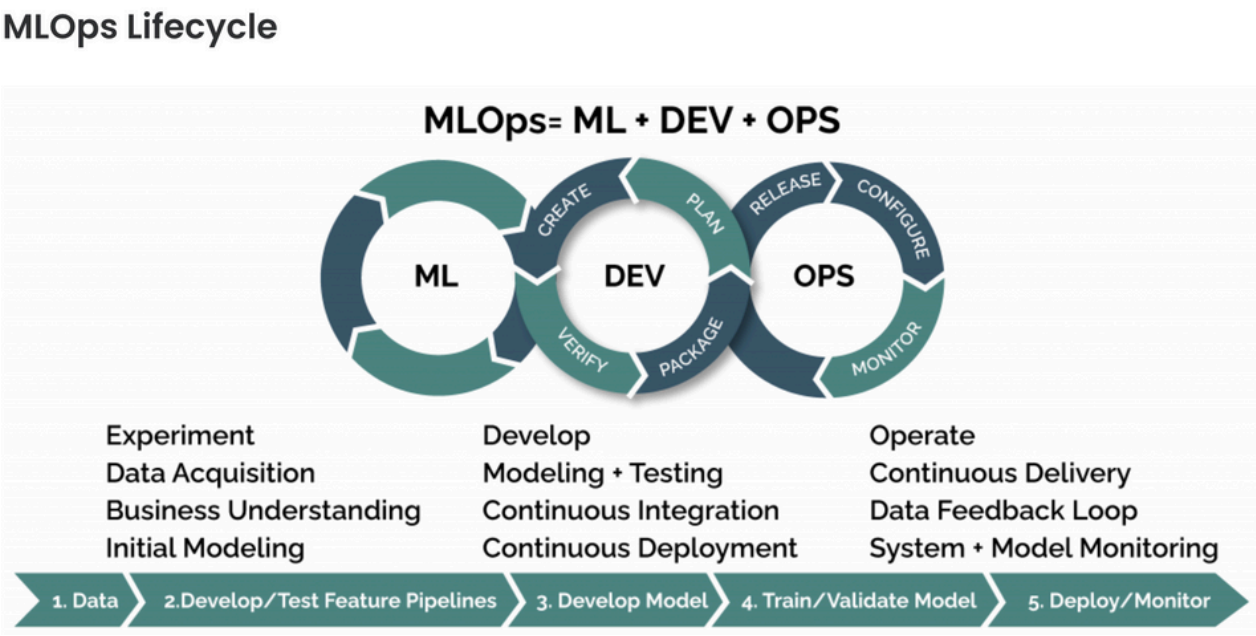


Figure 3: MLOps Lifecycle

Illustrating the relationship between the Machine Learning lifecycle, and Development operations to create a combined appraoch called Machine Learning Operations.

The role of MLOps as a catalyst for developing responsible AI through a design-thinking lens

Having defined RAI, discussed design thinking and described MLOps, this essay will now examine combining these approaches to provide a mechanism through which RAI can go beyond theoretical conclusions and become a tangible part of the development process. Developing a framework, such as responsible MLOps, allows ethical considerations to drive innovation throughout the entire machine learning lifecycle (Biswas et al.,2024). Moreover, it provides an approachable and practical route to compliance with upcoming EU AI regulations for teams who may feel overwhelmed. This allows them to infuse responsible AI methods across the AI lifecycle. Our aim should not be to treat ethics as an afterthought at the end of development or as a logically untenable principle at the beginning of the development process, but rather to incorporate it from the beginning to ensure safer, more trustworthy, and ethical AI products, and to identify any potential issues early in the lifecycle to avoid delay or the inability to deploy these products. This essay will now explain what this practically looks like by discussing how responsible techniques map across the entire ML lifecycle at three stages: 1. Design and preparation, 2. Development and evaluation and 3. Deployment, monitoring and operations.

Stage one: Design

In the design stage of MLOps, there are significant opportunities for setting a robust ethical foundation for the entire lifecycle, by using RAI principle-based tools, which will now be explored. This essay will not discuss the specificities of each tool or every tool that is available because this is out of scope of this essay. However, it will aim to discuss the overarching themes and take a light-touch approach to help companies start with one tool and add more to their arsenal as they get more comfortable. To clarify, this stage includes ideation, use case prioritisation, business understanding, data availability checks, data acquisition, and data preparation. Implementing a tool that supports all decisions made within each of these sub-stages and across the entire lifecycle requires the development of an ethical and responsible AI governance framework that goes beyond standard requirements concerning safety, privacy, and risk. It's common to have AI governance frameworks in MLOps, but this paper argues for one that is rooted, inspired and driven by AI ethics.

The International Standard Organisation (ISO) defines (ISO,2021) governance as "the system by which the whole organisation is directed, controlled, and held accountable to achieve its core purpose in the long run". A responsible design-thinking AI element in this context ensures that the governance framework extends beyond an internal perspective and influences external factors. It extends beyond a company's own value structures, including its teams, senior stakeholders, and customers. This encompasses the societal context in which they work, for instance, emerging AI regulations and collective expectations. A nuanced approach to governance allows organisations to achieve their objectives more effectively. Equally, the framework must be pragmatically oriented and provide enough flexibility to innovate to respond to the changing AI landscape. The purpose of this is not to water it down but to establish strong ethical foundations. In this way, multi-tiered governance frameworks can adapt to any complexities of the project. Integrating ethics into a team's governance framework provides them with clear mechanisms to demonstrate and document their consistent alignment with their values and principles.

A user-focused RAI governance framework would include actions around trustworthiness, fairness, accountability, transparency, explainability, and interoperability. These themes would expand beyond those commonly associated with governance, such as privacy, safety, and security. These themes still need to be included but enhanced and built upon. This proactive approach allows teams to mitigate problems as they arise and has clear pathways to do this. The OECD created a 'Catalogue of Tools & Metrics for Trustworthy AI' which contains many examples of responsible AI governance frameworks. This conserves time and financial resources for organisations.

Complementing this essay's argument to integrate RAI across the entire lifecycle, the Alan Turing Institute created a Process-Based Governance (PBG) Framework (Flynn et al., 2019) which allows exactly that. It is an architecture by which ethical practices are included at every point of the project lifecycle. It acts as a guiding star for teams to follow. It contains three levels which combine values, principles and processes. This acts as a practical and actionable mechanism to integrate responsible innovation across design, development and deployment. Organisations do not need to start from scratch and duplicate existing efforts as there are many open-source resources which can be used but, organisations need to tailor them to their relevant industry, user case and AI system so they can be as effective as possible (Tartaro et al.,2024).

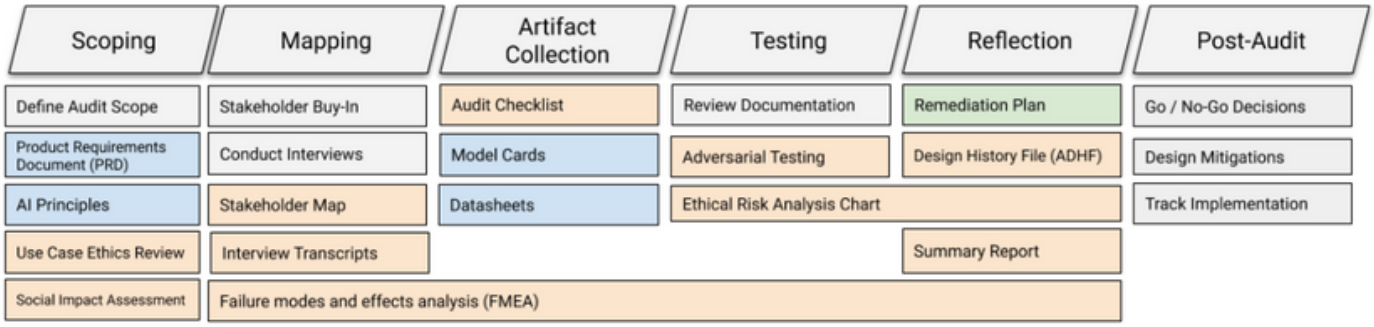
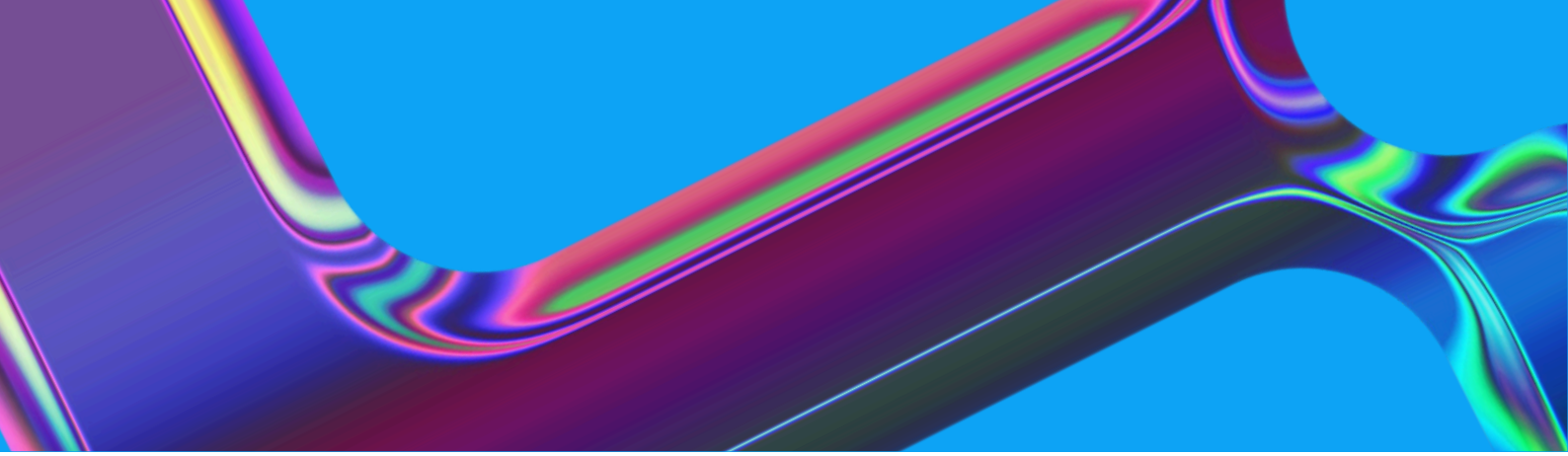


Figure 4 - Source: (Raji et al., 2020)
Diagram illustrating the key components of an Algorithmic Impact Assessment (AIA) framework



In conjunction, algorithmic impact assessments (AIA) is another RAI tool gaining traction that equally focuses on operationalising responsible AI principles and can be completed both in stages 1 and 2 (see Figure 4). This tool focuses on looking forward and partakes in a technique called horizon scanning in which cross-team stakeholders (i.e. data scientists, AI engineers, ethicists and security experts) use reflexive exercises to ascertain what potential impact the project they are proposing to build could have, and then what actionable early pro-active mitigation strategies they could use to prevent this. In agreement with Tartaro et al. (2024), it is vital that teams go beyond the standard AI risk assessment in the MLOps lifecycle and integrate it with the assessment of ethical dimensions, which the AIA tool fulfils.

An example of one recent tool (in the image above) is the SMACTR framework (which stands for Scoping, Mapping, Artefact Collection, Testing, and Reflection). This example demonstrates how algorithmic auditing can go beyond standard processes to apply ethical and societal considerations to the end-to-end development lifecycle. Within this model, a lot of documents are needed to fulfil the requirements such as AI principles documents, Use Case Ethics reviews etc. (Raji et al.,2020). To businesses, the beginning of their ethics journey is too heavy, which is why this paper offers a lighter touch to begin with, with which organisations can gain momentum and knowledge to then build themselves up to frameworks such as SMACTR.

Weaving RAI into the standard impact assessment provides a more holistic view as it facilitates a deeper comprehension of the "multidimensionality and context-sensitive AI risks" (Tartaro et al.,2024). Uncovering algorithmic risks within the context they would work in, goes beyond assessing risk within a vacuum and appreciates the environment, society and human context it would function in. This tool improves a project as key RAI principles such as fairness, transparency and non-maleficence are demonstrated through the discussion methods in this tool and a granular examination of all imaginable outcomes is assessed (Ashard et al.,2024). An example of an AIA in use is from the International Standards Organisation (ISO) under ISO 42005. This standard combines documentation from a plethora of areas across AI systems including internal policies, related risks, internal organisation, management guidance, data use, third-party relationships and others. Moreover, it assesses actual and potential impacts arising from system failure and misuse and aims to measure these to address any potential harms and amplify the benefits. As stated by ISO, this standard aims to document the impacts an AI service or product could have across the entire lifecycle, from ideation to deployment, towards individuals, groups of individuals or society.

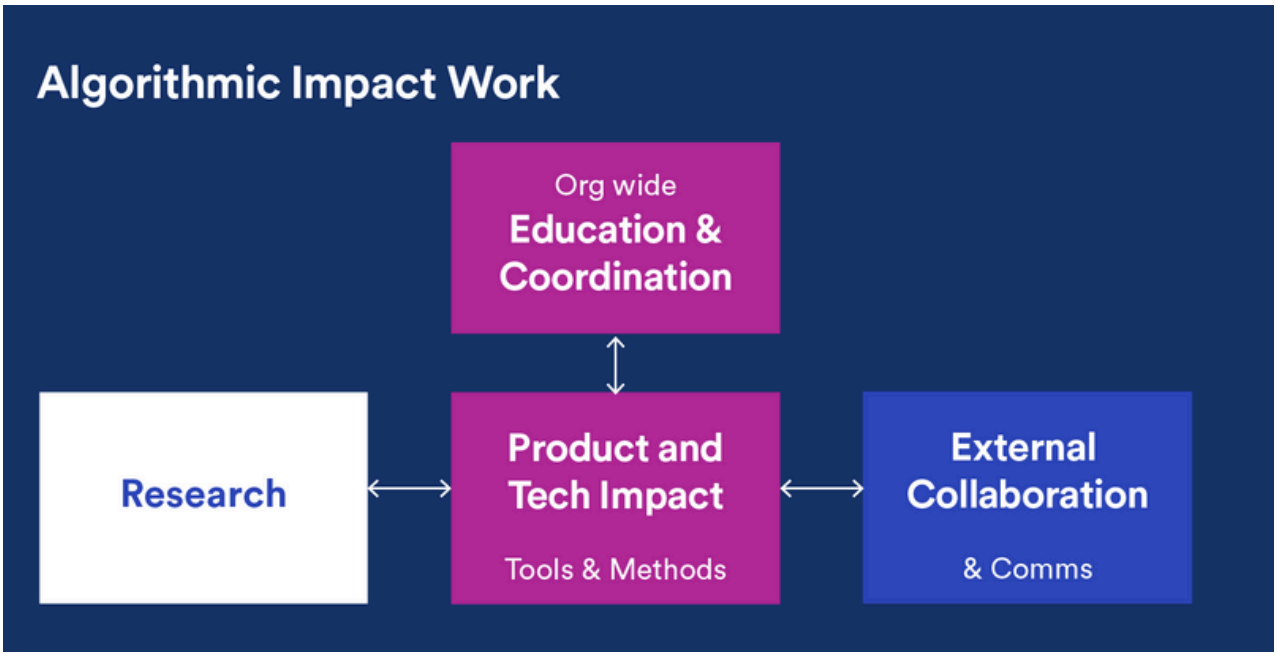


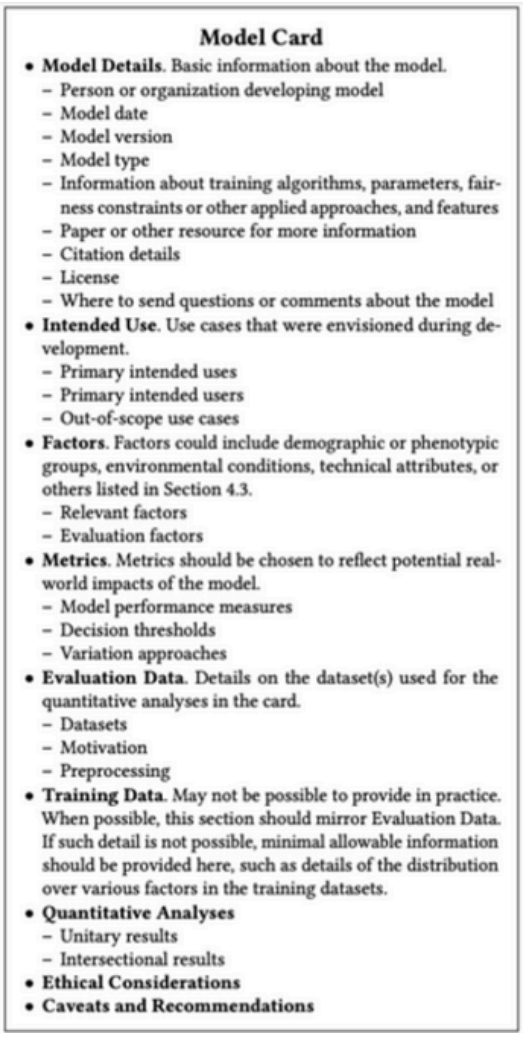
Figure 5 : Source - (Spotify, 2022)
Overview of Spotify's own Algorithmic Impact Assessment (AIA) they designed in 2022.

One case study from industry is of Spotify's use of an AIA to support their development of a safe platform serving millions of customers, and combines this with their goal of "taking responsibility for the impact we have on creators, listeners, and communities" (Spotify,2022). Spotify's AIA covered four areas including research, product and tech impact, external collaboration and internal education and coordination (see Figure 5). Through using this tool, they now have a team to assess and address unintended harmful outcomes, created an ecosystem across the company for advancing responsible recommendations and algorithmic systems, and lastly introduced governance, central guidance and best practices across approaches to personalisation, data usage and content recommendation. They state that this tool helped them "turn principles into practice" (Spotify,2022), supporting their goal to operationalise concepts into their day-to-day products and engineering practices. This process ensures that their responsibilities are not owned by only one team, but that they get both internal and external perspectives (Spotify safety advisory council) and that problem-solving these issues requires cross-discipline teams.

This is a relatively new approach and highlights the need for it to continue to evolve and iterate to deliver the best results, as currently there is a multitude of different formats and lack of standardisation. However, Ashar et. al (2024) in their research with over 107 practitioners building ML systems, found it an invaluable tool to advance a team's understanding of the potential harms of algorithmic systems and also found it uncovered mitigation mechanisms to prevent complications later. Having now explained the tool, discussed its success in an industry setting, highlighted its shortcomings, and proved its ability to operationalise RAI within the first stage of MLOps successfully. This essay will now explore opportunities and tools to be used in stage 2.

Stage two: Development and evaluation

Stage 2 of the MLOps lifecycle is building and evaluating. Responsible AI plays a pivotal role at this stage by providing tools to attempt to enhance fairness, which can be demonstrated to stakeholders, developers, and auditors. Utilizing RAI transparency methods to showcase these achievements boosts user confidence and trust, as the AI system becomes more comprehensible to both users and AI teams. Amongst the various transparency documentation methods, four common types are Model cards, Methods cards, AI cards and Use Case Cards. Each addresses RAI considerations slightly differently. This paper focuses on Model Cards, as they are specifically designed for AI practitioners who would be implementing these methods, and as this paper aims to approach RAI with a user-centred focus this seems the best tool.



Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

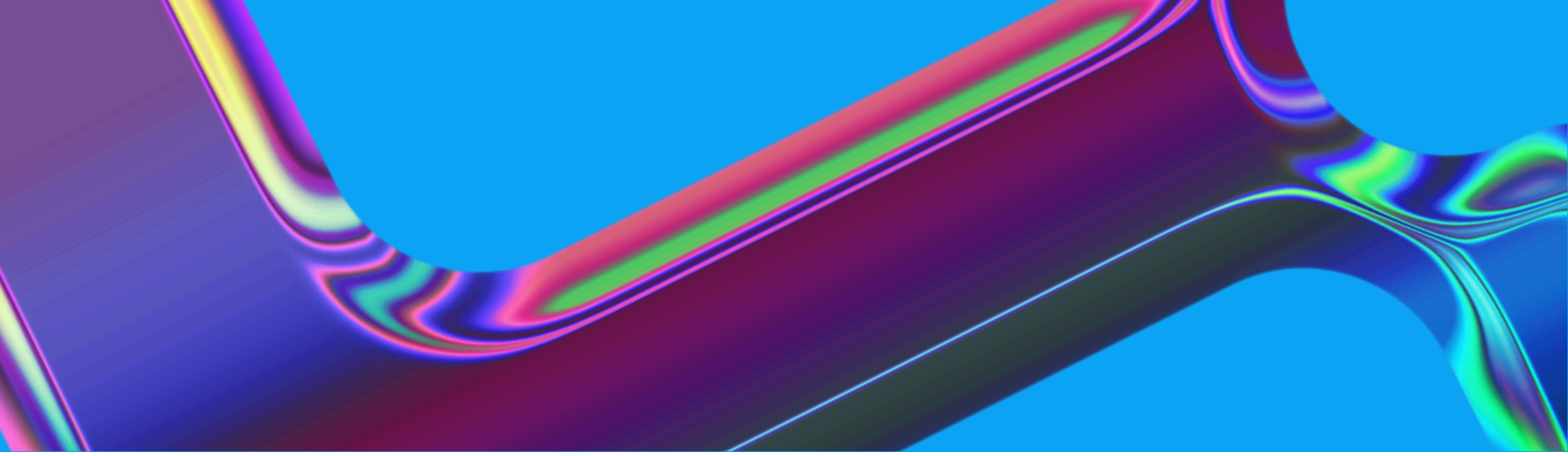
Figure 6 - Source: (Mitchell et al.,2019)
This image illustrates the key components of a Model Card, a documentation framework, which is designed to improve transparency in machine learning models.

Model cards emphasise metrics and benchmarking, particularly model performance measures, decision thresholds, approaches to uncertainty and variability and disaggregated evaluation of model performance unitary and intersectional factors (Mitchell et al.,2019). They adopt a mathematical approach to explainability, employing techniques such as saliency maps, path-integrated gradients and feature attribution. Although highly technical and structured, Model Cards are well-suited for use during stages 2 and 3, as they offer development teams the flexibility to include their insights, including findings from other RAI tools utilised earlier in the lifecycle. For these to be impactful teams need to provide clear information about a model's intended use, limitations, and potential impacts, Model Cards can then help ensure the AI system stays aligned with fairness, safety and transparency goals.

Model Cards can expose biases by documenting model performances across diverse populations, outlining appropriate use case, and identifying scenarios where the model should not be applied. Additionally, they address societal implications, fostering awareness of potential ethical concerns, such as privacy violations or discrimination. While various organisations have developed their versions, Google's model cards are among the most widely adopted, enabling developers to compile model information effectively (see Figure 6). Employing transparency documentation methods during this stage of the MLOps lifecycle demystifies the 'black box' nature of models, facilitates deeper interrogation, and proactively highlights potential risks.

Stage three: Deployment, monitoring and operations

In Stage 3, the focus of MLOps shifts to the deployment and monitoring of models. This stage presents various risks, with key challenges including model drift—where the model's performance deteriorates over time—unexpected behaviours in real-world scenarios that expose the system to adversarial attacks with unintended consequences, data quality issues impacting accuracy, and scalability problems where the deployed model fails to handle varying workloads reliably. To ensure accountability, bias mitigation, and robustness during this phase, the integration of human oversight is essential. While the previously discussed tools provide valuable support for addressing potential risks, human judgement remains critical in responding to real-time, evolving challenges that arise during deployment and monitoring. Employing an iterative feedback process that balances human, and machine inputs is therefore crucial. A Human-in-the-Loop (HITL) approach can enhance performance, and accuracy, and mitigate negative societal impacts.



HITL refers to the process of training models through semi-supervised learning, in which human feedback is combined with a small set of manually labelled data and a larger set of unlabelled data to enable automatic machine annotation. This approach enhances the model's ability to iteratively label data, accelerate development cycles, and embed standardised ethical AI practices into the model. However, it is recognised that humans bring their own biases to the annotation process, making it necessary to guard against ethical distortions stemming from the annotator's values. For this reason, establishing a governance process at the outset of the cycle is vital. A code of ethics, defined during this stage, can guide annotators in making fair judgements aligned with pre-agreed principles, thereby making the system's ethical decisions more transparent and consistent.

HITL underscores the inherent tension between transparency and efficiency in AI systems. While it does not promise perfection, it seeks to manage potential risks while maintaining ultimate human control over the system. Combining the strengths of human and machine annotation compensates for the limitations of relying solely on either (Chen et al., 2023). As Solar-Lezama from MIT notes, "Any decision that is important should not be made by a [language] model on its own" (Wallach and Allen, 2008). A qualified human must be involved to prevent biases and errors, ensuring the system remains responsible and meets user needs. Although HITL cannot guarantee flawless AI systems, it provides a strategy within the MLOps lifecycle to mitigate risks, integrate ethical principles, and ensure responsible deployment and monitoring.

Through the inclusion of human oversight, HITL addresses risks such as AI hallucinations, model drift, ethical bias, and adversarial attacks. It also highlights the importance of a collaborative human-machine dynamic grounded in an iterative and adaptable ethical governance framework, aligning human decision-making with technical robustness and shared values. By fostering trustworthiness and fairness, HITL enhances user confidence in AI systems, ensuring their safety is prioritised. Ultimately, this approach allows AI systems to adapt to evolving challenges while safeguarding user interests.

Conclusion

This paper explores the necessity of grounding RAI by transitioning from abstract ethical principles to concrete actions embedded within the MLOps lifecycle. Bridging the gap between ethical guidelines and their practical application within a technical workflow overcomes trust deficits in AI systems, mitigates potential harms, and unlocks opportunities for positive societal impact. By integrating RAI into MLOps, teams achieve greater cohesion as diverse roles collaborate to build a holistic understanding of the AI system. This integration is not merely an ethical imperative but a strategic advantage.

Embracing a user-centred design-thinking approach allows organisations to shift from reactively addressing problems to proactively delivering robust, real-world AI solutions. Such solutions avoid costly last-minute fixes to fairness or bias issues at deployment and prevent brand reputational damage arising from these failures. This proactive stance, coupled with practical tools such as AIA, model cards, HITL techniques, and flexible governance structures, fosters a deeper and more nuanced identification of potential risks and allows mitigation strategies to be built early in the process. These tools operationalise RAI principles by involving diverse stakeholders in reflexive exercises that move beyond the narrow, technical focus of risk assessments to also encompass societal and ethical dimensions. For example, Spotify's implementation of RAI demonstrates how principles can be effectively translated into practice (Ashar, 2024), fostering a culture of shared responsibility across team members rather than burdening individual members.

The sustainable and safe adoption of AI systems will hinge on the creation of systems that not only prioritise productivity and efficiency but also align with RAI goals. Integrating RAI principles into MLOps through a design-thinking lens provides a practical pathway for technical teams to achieve this balance. Organisations committed to fairness, accountability, and transparency, position themselves as proactive leaders, equipped to navigate evolving regulations while differentiating themselves in the market as pioneers of safe, trustworthy AI products. This approach paves the way for a future of widespread adoption of AI systems and products that truly serve user needs and foster innovation whilst safeguarding against potential harm.

Bibliography

Aragon, C., Guha, S., Kogan, M., Muller, M. and Neff, G., 2022. Human-Centred Data Science: An Introduction. Cambridge, MA: MIT Press.

Ashar, A., Ginena, K., Cipollone, M., Barreto, R., & Cramer, H. (2024). Algorithmic Impact Assessments at Scale: Practitioners' Challenges and Needs. Journal of Online Trust and Safety, 2(4). https://doi.org/10.54501/jots.v2i4.206

Ashard, A., Ginena, K., Cipollone, M., Barreto, R. and Cramer, H., 2024. Algorithmic Impact Assessments at Scale: Practitioners' Challenges and Needs. Journal of Online Trust and Safety, 2(4). Available at: https://doi.org/10.54501/jots.v2i4.206 [Accessed 25 Oct. 2024].

Billeter, Y., Denzel, P., Chavarriaga, R., Forster, O., Schilling, F., Brunner, S., Frischknecht-Gruber, C., Reif, M. and Weng, J., 2024. MLOps as Enabler of Trustworthy AI. 2024 11th IEEE Swiss Conference on Data Science (SDS), pp. 37–40. Available at: https://doi.org/10.1109/SDS60720.2024.00013.

Biswas, D., Chakraborty, D. and Mitra, S., 2024. Responsible LLMOps: Integrating Responsible AI Practices into LLMOps. Webit Sofia Edition. Available at: https://blog.webit.org/2024/09/18/thrilled-announce-debmalya-biswas-director-data-analytics-ai-wipro/ [Accessed 15 Oct. 2024].

Brown, T., 2009. Change by Design: How Design Thinking Creates New Alternatives for Business and Society. HarperCollins.

Chen, X., Wang, X. and Qu, Y., 2023. Constructing Ethical AI Based on the "Human-in-the-Loop" System. Systems, 11(11), p. 548.

Davenport, T.H. and Ronanki, R., 2018. Artificial Intelligence for the Real World. Harvard Business Review, 96(1), pp. 108–116.

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., de Prado, M.L., Herrera-Viedma, E. and Herrera, F., 2023. Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation. Information Fusion, 99, p. 101896.

Dorst, K., 2011. The Core of 'Design Thinking' and Its Application. Design Studies, 32(6), pp. 521–532.

European Parliament, 2024. Resolution on Artificial Intelligence and Its Implications for the Economy and Society. [online] Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf [Accessed 25 Sep. 2024].

European Parliament and Council, 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a European Approach for Artificial Intelligence. [online] Available at: https://eur-lex.europa.eu/eli/reg/2022/2065/oj [Accessed 25 Sep. 2024].

Bibliography

International Organization for Standardisation (ISO), 2024. Building a Responsible AI: How to Manage the AI Ethics Debate. Available at: https://www.iso.org/artificial-intelligence/responsible-ai-ethics#toc7 [Accessed 12 Aug. 2024].

Jobin, A., Ienca, M. and Vayena, E., 2019. The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence, 1(9), pp. 389–399.

International Organisation for Standardisation (ISO), 2021. ISO 37000:2021 Governance of Organizations — Guidance. Geneva: International Organization for Standardization. Available at: https://committee.iso.org/ISO_37000_Governance [Accessed 20 Nov. 2024].

(Treve, M., Omont, N., Stenac, C., Lefevre, K., Phan, D., Zentici, J., Lavoillotte, A., Miyazaki, M. and Heidmann, L., 2020. Introducing MLOps. O'Reilly Media.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T., 2019. Model Cards for Model Reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 220–229.

Paton, B. and Dorst, K., 2011. Briefing and Reframing: A Situated Practice. Design Studies, 32(6), pp. 573–587.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F. and Dennison, D., 2015. Hidden Technical Debt in Machine Learning Systems. In: Advances in Neural Information Processing Systems 28 (NIPS 2015). Available at: https://papers.nips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf [Accessed 6 Aug. 2024].

Tamburri, D.A., 2020. Sustainable MLOps: Trends and Challenges. In: 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC). IEEE, pp. 17–23.

Tartaro, A., Panai, E. and Cocchiaro, M.Z., 2024. Risk Assessment Using Ethical Dimensions. In: Braunschweig, B. et al., eds. AITA: AI Trustworthiness Assessment. AI Ethics, 4, pp. 1–3. Available at: https://doi.org/10.1007/s43681-023-00397-z.

UK Government, 2023. Regulation of AI (Human Rights) Act 2023. [online] Available at: https://www.legislation.gov.uk/ukpga/2023/50 [Accessed 25 Sep. 2024].
Wallach, W. and Allen, C., 2008. Moral Machines: Teaching Robots Right from Wrong. Oxford University Press.

World Economic Forum, 2023. Responsible AI Requires More than Technical Solutions. [online] Available at: https://www.weforum.org/agenda/2023/10/responsible-ai-requires-more-than-technical-solutions/ [Accessed 25 Sep. 2024].

Zhu, L., Xu, X., Lu, Q., Governatori, G. and Whittle, J., 2022. Operationalising Responsible AI in Humanity-Driven Ecosystems. In: Responsibility and AI in Humanity-Driven Ecosystems. Springer, pp. 15–33.

Be a part of the BridgeAI revolution

Register and receive updates on events, programme developments and upcoming competitions.

bridgeai.net