EBOOK

Build Responsible Generative AI Applications: Introducing the RAFT Framework

Introduction: Great Power, Great Responsibility

Recent excitement around Generative AI — in particular Large Language Models (LLMs) — means organizations are pushing forward with the use of AI at an unprecedented pace. There has arguably never been a more pivotal time in the history of AI.

At the same time, it's important to stay grounded. The truth is that flaws within AI systems and the data they are built on can present — and have presented, even before the rise of Generative AI — real risks.

More than ever before, organizations need to think about building AI systems in a responsible and governed manner.

In this ebook, we will deep dive into those aforementioned risks plus introduce the RAFT (Reliable, Accountable, Fair, and Transparent) framework for Responsible AI, showing how it can be applied to both traditional and Generative AI systems.

Understanding Values for Responsible AI

Whether your business is scaling the use of AI across the organization, interested in experimenting with the latest developments in Generative AI, or simply looking to make sense of forthcoming regulation, it's important to have a set of guardrails defined for the use of AI at your organization.

Various standards organizations and governments have proposed frameworks for AI values — see Table 1. These are a good starting point, but in the next section, we'll take it one step further with a more specific, robust, and tested framework.

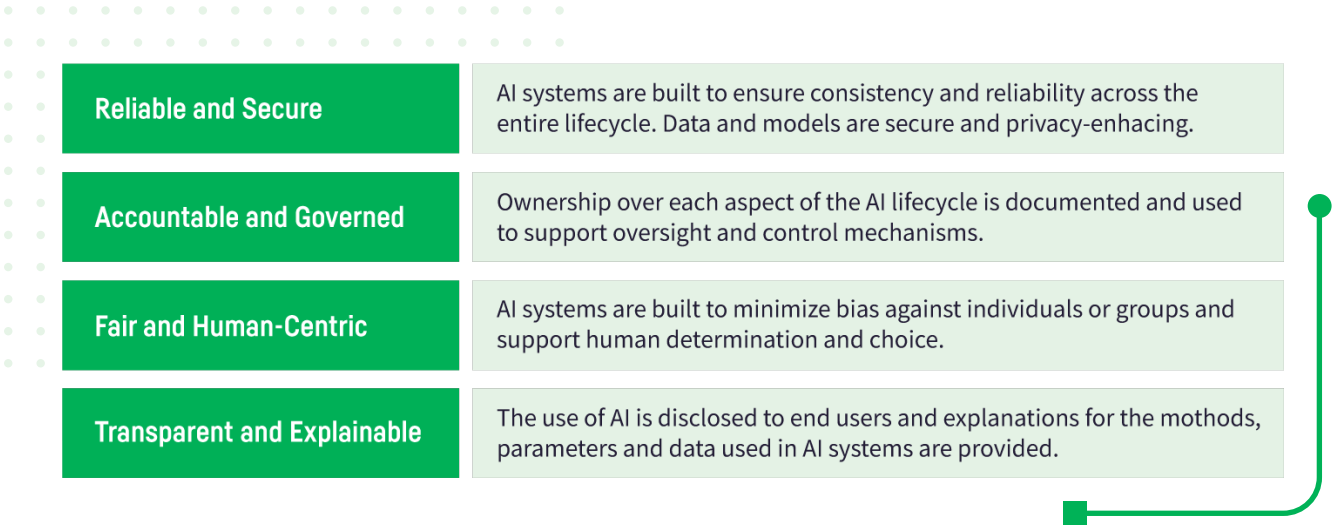
TABLE 1: MAPPING OF AI RMF TAXONOMY TO AI POLICY DOCUMENTS

AI RMF	OECD AI Recommendation	EU AI Act (Proposed)	EO 13960
Valid and reliable	Robustness	Technical Robustness	Purposeful and performance driven Accurate, reliable, and effective Regularly monitored
Safe	Safety	Safety	Safe
Fair and bias is managed	Human-centered values and fairness	Non-discrimination Diversity and fairness Data governance	Lawful and respectful of our Nation's values
Secure and Resilient	Security	Security & Resilience	Secure & Resilient
Transparent and accountable	Transparency and responsible disclosure Accountability	Transparency Accountability Human Agency and oversight	Transparent Accountable Lawful and respectful of our Nation's values Responsible and traceable Regularly monitored
Explainable and Interpretable	Explainability		Understandable by subject matter experts, users, and other, as appropiate
Privacy-enhanced	Human values; Respect for human rights	Privacy Data governance	Lawful and respectful of our Nation's values

Source : NIST AI RMF V2

The RAFT Framework for Responsible AI

To support the scale of safe AI, Dataiku offers a number of built-in features for responsible design, deployment, and governance. These features align with a value framework similar to those proposed in Table 1, which any organization can use as a starting point for their own Responsible AI programs.



Note: See page 11 for the full framework.

These values make up Dataiku's baseline approach to Responsible AI — we call it RAFT for Reliable, Accountable, Fair, and Transparent. The values outlined in the RAFT framework are crucial for the development of AI and analytics, and they cover both traditional and new methods in Generative AI.

To effectively execute on these principles requires understanding the potential risks and impacts of technology. In the rest of this ebook, we will cover the specific risks of Generative AI and broader approaches to Responsible AI practices.

Our understanding of the risks and necessary steps to reduce potential harm from AI is meant as a starting point for organizations looking to build best practices in governed and Responsible AI. In addition to building best practices, these guidelines can support organizations in their readiness efforts toward upcoming regulations, such as the EU AI Act.

As the field of AI continuously evolves, so do our approaches to safe and scaled uses of new technology. We encourage readers to take the suggestions and recommendations here and adapt and expand them as needed per industry and regulatory standards.

Risks to Using Generative AI in the Enterprise

AI systems are inherently socio-technical, which means "they are influenced by societal dynamics and human behavior."[1] Acknowledging the way AI impacts society and vice versa allows us to better anticipate potential negative consequences and proactively reduce or prevent those harms before they occur.

In addition to the socio-technical risks from Generative AI, there are also emerging legal considerations around privacy and copyright infringements. Below, we list a number of risks that may arise in the use of Generative AI in the enterprise. These risks are common across various types of Generative AI technology but will surface in different ways across use cases:

- **Toxicity:** Toxic, obscene, or otherwise inappropriate outputs.
- **Polarity:** Unfair positive or negative attitudes to certain individuals or groups.
- **Discrimination:** Model performance is less robust for certain social groups.
- **Human-Computer Interactions[2]:** Over-reliance on the outputs of AI due to perceived sentience or blind trust in an automated system.
- **Disinformation:** Presenting factually incorrect answers or information.
- **Data Privacy:** Input data shared back to 3rd-party model providers and possibly shared as future outputs to non-authorized users.
- **Model Security:** Ability for a user to circumvent security protocols intended to prevent social-technical harms or gain access to unauthorized data.
- **Copyright Infringements:** Redistribution of copyrighted material, presented as original content.

1. https://www.nist.gov/news-events/news/2023/01/nist-risk-management-framework-aims-improve-trustworthiness-artificial#

The potential harms listed here are not exclusive to language models, but they are heightened by the use of natural language processing (NLP) techniques to analyze, categorize, or generate text in a variety of business contexts.

Understanding and addressing these risks before implementing an LLM or other Generative AI techniques into an AI system is crucial to ensure the responsible and governed use of the latest technology. By thinking through potential harms, designers, developers, and deployers of LLMs can set thresholds for risk and incorporate mitigation strategies in each stage of the AI lifecycle.

Risks of Generative AI by Context

Beyond the inherent, high-level risks associated with Generative AI technology, businesses should consider the context in which the system will be deployed.

A baseline approach is to assess the use case across two dimensions:

1. Target of analysis, which focuses on the type of data or documents that the model will make use of to generate output. Corporate or business documents include items as invoices, legal agreements, or system documentation. Individual or personal data can include traditional tabular information about a person's specific characteristics, as well as call center transcriptions or text written by an end user. Academic or domain-specific texts are typically used in industry research and analysis, such as medical publications, manufacturing research, or legal codes.

2. Delivery method, which looks at how the output of a model is delivered to end users. In the first instance, the output is shared as a report, recommendation, or suggested actions in response to a single query. This is different from the next category of virtual assistants, chatbots, etc., that respond in a human-like way to end users' queries. The final category is automated processes, such as sending mass marketing emails or robotic process automation (RPA).

Each category within these two dimensions will carry different risk tradeoffs and strategies to prevent harm to the business, clients, and broader society.

Target of Analysis	Delivery Method
Corporate or Business Documents While social risks from the use of these documents are lower, it is important to ensure the documents are up to date and relevant for the question at hand. This should include designing the input parsing strategy to leverage the correct documents and content to answer a given query. Prevent leakage of sensitive or copyright material into the model output — users should not be able to circumvent model parameters to	**Output Shared as a Report, Recommendation, or Suggestion Actions** Generally this type of delivery ensures that end users have control over how the output is shared or deployed down the line. It provides an opportunity for review of outputs before action is taken, allowing the end user a chance to review explanations and provide feedback.
Individual or Personal Data Personal data on customers, users, or patients should be protected from unauthorized access. Models that make use of personal information to draw conclusions about individuals should be tested for fairness across subpopulations of interest, language output should not be toxic or reinforce stereotypes about groups of people, and end users should know that their data is being used to train and deploy models (with the ability to opt out from this usage). Take care that personal information or details that are not relevant to the specific use case are not used in the model or shared	**Virtual Assistant** Virtual assistant delivery includes chatbots, text to voice responses, and question answering. These interactions should be clearly marked as a computer/model and not a human agent. Extra precautions should be taken to make it clear the model is not a sentient agent and could potentially provide incorrect responses. Additionally, generative text models should be developed with guardrails around the type of language and discussion permissible by the model and end users to prevent toxic or harmful conversation from occurring.
Academic or Domain-Specific Texts Texts should be carefully cultivated to ensure fit with the intended use case. Model outputs should be able to provide citations to real texts in the corpus to support any generated answers or recommendations.	**Automated Process** Results of a Generative AI model may be passed directly to an end user without review. Automated processes are typically used to scale AI use, meaning real-time human intervention is not possible. Regular review of model outputs, quality, and the ability to pause automated processes when necessary is critical to catch any potential harms. Ensure you clearly document accountability over stopping an automated

Concerns Based on Expected Audience

One important dimension of Generative AI use cases we have not yet covered is the type of audience or expected reach for model outputs. The audience for model outputs are usually business users, consumers, individuals or some combination of both groups.

However, no matter the expected audience for model outputs, there are core criteria that should be met in the deployment of any AI system — these four criteria further the goals of reliability and transparency as detailed in the RAFT principles and support broader trust in AI systems:

#1 Feedback Mechanism

End users should have the ability to provide feedback on the quality of output generated by a model as either free text or a quality score. This feedback should be collected alongside the user's prompt and the model output, and it should be regularly reviewed by the model owners in order to adjust parameters and ensure outputs are consistent with expectations.

#2 Transparency on Use of AI for Output

When interacting with an AI system — such as through prompts, automated processes, or generated text — clear and visible messaging should indicate to an end user that they are receiving output from a model and not a human agent. Any real-time interaction with a virtual assistant should be built with guardrails to prevent off-script responses or the perception of AI sentience.

#3 Explanations for Given Response

Where possible, explanations for model outputs must be provided to end users. For non-Generative AI tasks, individual record explanations can be computed using ICE or SHAP methods. Text or output from Generative AI can provide explanations by displaying references to the specific text or underlying analysis used to form the response to a query.

#4 Education on Limitations of the Model

Users should have access to a basic understanding of how models produce predictions or generate output. In addition, limitations of the model should be made clear. Limitations can include the relevancy of data used to train the model and the overall consistency of answers produced by the model. Best practices for prompt engineering and the expected use for model outputs should be documented for users.

Assessing Potential Impacts of AI

Before deploying an AI system —generative or not — it's important to assess the potential impact on individuals or groups once a solution is in use. While specific impacts or unintended consequences will vary from use case to use case, Dataiku suggests two broad dimensions that can be used to understand impact from a deployed AI system. These impacts and potential risks are based on standards such as the NIST Risk Framework or the EU AI Act and are meant to guide our customers as they implement AI systems.

Depending on the use case, an AI pipeline may have more than one type of model providing output. For example, the next best offer project[3] uses both a traditional recommendation model and an LLM to help write the text of an email to a customer.

In such an instance, it is important to assess impact and apply Responsible AI principles to both sets of models in the project. Potential bias or poor performance from a recommendation model will have different impacts than bias or toxicity from a language model.

The risk scoring for unintended consequences is based on two variables:

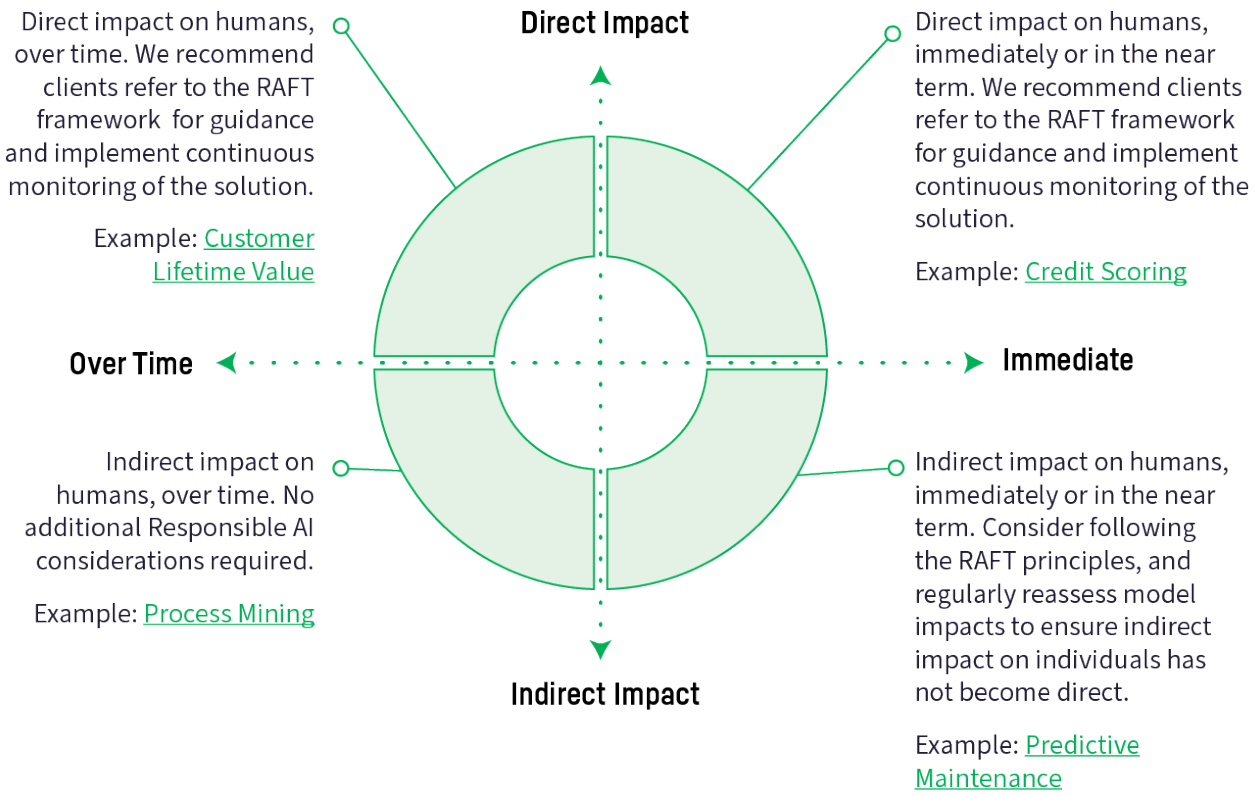
1. Whether the risk could materialize as a harm to individuals and groups directly because of the solution's implementation or indirectly because of some constellation of factors that are difficult to qualify at the time of deployment.

2. Whether the risk could materialize as a harm immediately or over a longer period of time.

This results in two larger guiding questions about the nature of an AI system :

Q1 : Does the output of this project lead to direct impact on individuals or groups?
Q2 : Is the impact felt immediately or observed over time?

Putting these two variables together, we qualify the solution's Responsible AI considerations in one of the following categories :



Putting Principles into Practice

How do we move from defining principles, potential risks, and impacts of AI to implementing best practices and guardrails in the development of AI systems?

Let's return to the RAFT framework, which provides a baseline set of values for safe AI and that can serve as a starting point for your organization's own indicators for Responsible AI. We encourage the governance, ethics and compliance teams at your organization to adapt the framework to accommodate specific industry requirements, local regulations or additional business values.
As with assessing impact, it is necessary to apply the principles from the RAFT framework to all models (both traditional and generative) used in an AI system.

#1 Reliable and Secure

Privacy and Security

- Document and anonymize personal data according to regulations.
- Regularly check model outputs for copyrighted or sensitive data.
- Ensure models can not be misappropriated or altered by malicious actors.

Model Quality and Robustness

- Integrate unit tests and debugging into the model build process.
- Test models for robustness from adversarial attacks, data, or concept drift.
- Evaluate models on real-world datasets before deployment and confirm attainment of performance metrics.
- Document (and use) relevant benchmarks for model evaluation.

Usage Monitoring and Assessments

- Clearly document requirements and specifics of when and how a model or output can be used.
- Continuously monitor production pipelines to ensure deployment is consistent with intended usage defined by the business.
- Collect and analyze model outputs on a regular basis to ensure model consistency on key metrics.
- Regularly review and assess generative model user prompts to ensure usage remains aligned with intent.

#2 Accountable and Governed

Documentation and Ownership

- Document relevant RAFT checks and decision points in project wikis and appropriate governance workflows.
- Ensure clear ownership for each stage of the development cycle as well as corporate accountability for potential failures.

Third-Party Model Governance

- Document all instances of third-party models in use across systems.
- Provide best practices and limitations to end users interacting with third party models.
- Review and authorize third-party models in accordance with governance or responsible AI policies where possible.

Oversight and Sign-Off

- Assign roles and requirements for each stage of the AI pipeline, cross-checked against the RAFT framework and the full development cycle.
- Require sign-off from relevant parties before moving to the next stage of pipeline build.

#3 Fair and Human-Centered

Bias Measurements

- Document potential sensitive attributes (SAs) in datasets.
- Measure and document disparate impact of SAs across outcome variables and relevant subpopulations.
- Check for proxy variables against SAs in raw data and engineered features.
- Assess dataset features for potential human bias in entry or encoding practices.
- Employ data bias mitigation as required (i.e., remove proxy variables, weight data, or remove encodings).
- Check system design and data collection practices for automation, sampling, or confirmation biases.

Thresholds and Acceptable Deviations

- Determine suitable, use case-specific fairness metrics prior to model building.
- Assess risk/value of deviation from the selected fairness metric(s), and document acceptable risks of deviation.
- Reevaluate metrics and thresholds after model build and before final deployment.
- Provide monitoring teams with guidance for preventative risk monitoring.

Impact and Unintended Consequences

- Gather relevant stakeholders prior to development to map out expected impacts and potential unintended consequences.
- Measure model fairness in accordance with set guidelines when pipeline is in production.
- Reevaluate new data for disparate impact across sensitive attributes and new potential proxy relationships on a consistent basis.

#4 Transparent and Explainable

Data Lineage and Traceability

- Document all datasets used as foundations for models according to principles outlined in Datasheets for Datasets[1], including how the data was collected, whether it is representative of the population of interest, and with what intent it was collected.
- Document the rationale for using specific datasets and how new features should be used in downstream/alternative pipelines.
- Document the rationale for and steps taken toward data cleansing, transformation, or other feature engineering.

Explainability and Interpretability

- Provide explanations (where possible) for new predictions from all deployed models.
- Build dashboards that contextualize individual predictions against all training data and overall feature importance of the selected model. Share these dashboards with end users or those affected by the model.
- Document why a given model was selected before deployment, including rationale for any custom evaluation metrics built into the model.

Reporting and Enablement

- Develop reporting that provides full documentation of the AI pipeline, all relevant decisions taken during build, and steps taken as part of the responsible AI framework.
- Provide clear guidelines on the use and intended purposes of the AI system, as well as those use cases for which it should not be employed.
- Provide a mechanism for recourse or feedback if an end user is not satisfied with outcomes, and review this feedback in a consistent manner.
- Clearly state when AI is used to produce outputs to end users, consumers, or other affected parties.

Note: Specific methods to assess and manage the bias of language models or other Generative AI are still in development. When building or fine tuning a model, developers should use diverse and representative datasets and check model performance against risks like polarity,

Conclusion: An Evolving World & Ongoing Considerations

It's only the beginning for Generative AI, which means we're only at the beginning of our understanding of the extent of the opportunity — as well as the risks — it presents.

In addition to the RAFT guidelines for Responsible AI, Dataiku is proud to offer a comprehensive training on how to implement Responsible AI in practice. These courses are available for anyone who wishes to gain hands-on experience with measuring bias, building fairer models, and creating compelling reporting on data science pipelines with Dataiku. The training covers core concepts in Responsible AI through the lens of classification models, but much of the teachings around bias, explainability and reliability can be applied to Generative AI models as well.

Access the Responsible AI training on Dataiku Academy

Here are some additional considerations to keep top of mind moving forward when it comes to leveraging Generative AI in the enterprise.

Third-Party Models

In many cases today, businesses will rely entirely on pre-trained, third-party models to support a Generative AI use case. In this situation, risks arise from the inability to know how the model was trained, not having confirmation that it is not biased or unreliable, and having no control over how input data is shared back to the third-party provider for retraining the model.

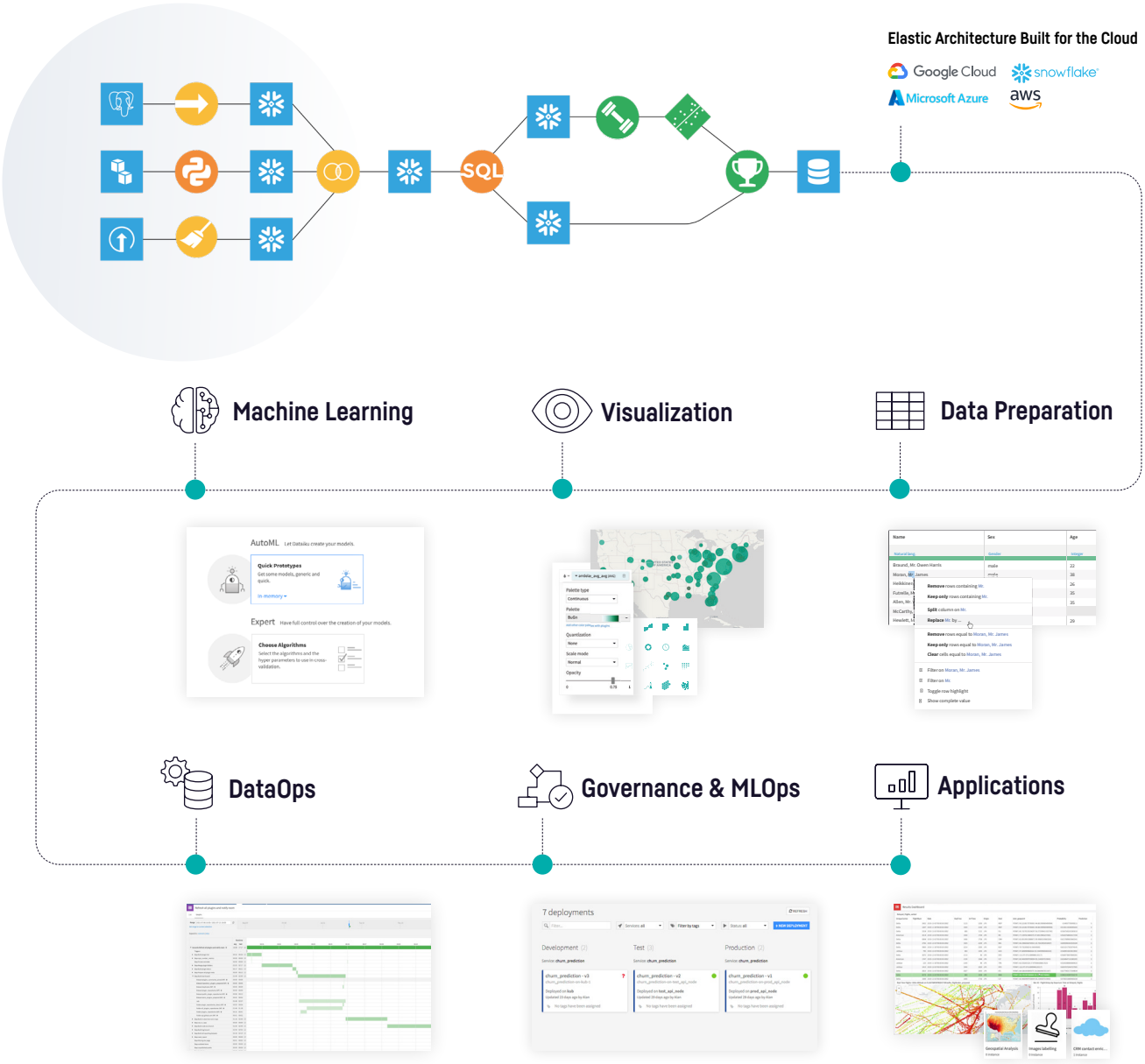
The use of third-party models should be closely monitored across the organization, and prompts should be regularly reviewed to ensure private data is not supplied as inputs to the model. Additionally, education on how to work with the third-party model (plus test the reliability of outputs) should be available for users.

Environmental Impacts

An additional area of concern is that of the financial and environmental cost of training and deploying LLMs. Given the numerous pre-trained models available for commercial and open-source use, many organizations do not need to build an LLM from scratch. Instead, many businesses opt to fine tune existing models on specific data, which is far less costly and requires less compute resources.

Fine tuning an existing model is therefore a benefit to businesses who wish to avoid the pitfalls from using third-party models and reduce the cost of training a new model. However, it is important to note that using a generative model in a production environment requires the use of GPUs, which will incur higher computational and environmental costs.

Everyday AI, Extraordinary People



Dataiku is the platform for Everyday AI, enabling data experts and domain experts to work together to build AI into their daily operations. Together, they design, develop and deploy new AI capabilities, at all scales and in all industries.

Authors

JACOB BESWICK

Jacob Beswick is Dataiku's Director for AI Governance Solutions and Responsible AI. Jacob leads a team that works with organizations globally to develop their approaches to AI and analytics governance, ensuring that their objectives — for example, compliance, risk management, or responsible development and use — can be embedded in Dataiku's platform. Prior to joining Dataiku, Jacob served in the UK Civil Service, where he led a portfolio covering AI adoption, governance and regulation. As part of this, he represented the UK at the European Commission, drafted what has become the UK's proposals on AI regulation, and created BridgeAI, which is now being delivered by Innovate UK.

TRIVENI GANDHI

Triveni is a Jill-of-all-trades data scientist, thought leader, and advocate for the responsible use of AI who likes to find simple solutions to complicated problems. As Responsible AI Lead at Dataiku, she builds and implements custom solutions to support the responsible and safe scaling of AI. This includes hands-on training to put RAI principles into practice as well as change management programs that help business leaders translate ethical values into actionable indicators. Triveni also supports field product teams with innovative methods to govern and oversee complex AI problems — often serving as a technical bridge between subject matter experts and engineers. Triveni is the winner of VentureBeats Women in AI Awards in the category of Responsibility and Ethics in AI and was listed as 100 Brilliant Women in AI Ethics in 2022. She holds a Ph.D in Political Science from Cornell University.