Responsible Data Science Policy

Design and Implementation Guide

**Table of Contents**

[**1. Introduction**](#_heading=h.1fob9te) **1**

[1.1 PURPOSE AND INTENDED AUDIENCE](#_heading=h.3znysh7) 1

[1.2 ORGANIZATION](#_heading=h.2et92p0) 2

[1.3 UPDATES AND ASSISTANCE](#_heading=h.tqkkb5h2ko6f) 2

[**2. Policy Documents**](#_heading=h.tyjcwt) **3**

[2.1 RESPONSIBLE DATA SCIENCE POLICY](#_heading=h.3dy6vkm) 3

[2.2 PARENT PROCEDURE TEMPLATE](#_heading=h.1t3h5sf) 5

[2.3 ADJUDICATIVE PROCEDURE TEMPLATE](#_heading=h.4d34og8) 5

[2.4 PRESCRIPTIVE PROCEDURE TEMPLATE](#_heading=h.2s8eyo1) 7

[2.5 CONCEPTS AND TECHNIQUES INVENTORY](#_heading=h.17dp8vu) 10

[**3. Sample Artifacts**](#_heading=h.3rdcrjn) **10**

[3.1 POLICY EXCEPTION LOG](#_heading=h.26in1rg) 10

[3.2 DATA SCIENCE PROPOSAL FORM](#_heading=h.lnxbz9) 11

[3.3 DATA SCIENCE REVIEW FORM](#_heading=h.35nkun2) 11

[3.4 DATA SCIENCE REVIEW LOG](#_heading=h.1ksv4uv) 11

[3.5 DATA SCIENCE RELEASE FORM](#_heading=h.44sinio) 11

# 1. Introduction

## 1.1 PURPOSE AND INTENDED AUDIENCE

This document is intended to help compliance and legal professionals use the Licensio Responsible Data Science Policy Framework, a set of flexible template Policies and Procedures that support a wide range of maturity levels and sectors.

This framework has been developed to respond to a perceived market need for better internal risk management and external trust-building, especially as data science topics emerge in conversations with customers, regulators, investors, or the broader public.

The framework is designed around principles, listed in the policy, that have been chosen to produce technical, legal, and ethical results that both internal and external stakeholders can agree to. As such, this framework is not the most legally conservative or financially aggressive; it makes a number of assumptions and compromises that organizations should carefully consider prior to adoption.

Furthermore, this framework is not intended to provide sufficient controls to meet any specific law or regulation. While many of the elements of this policy are necessary for compliance with laws such as GDPR or CCPA or standards such as SOC2/TSC, this policy alone is not sufficient for such compliance. Organizations should carefully consider how this policy fits into their overall legal and contractual requirements prior to adoption, ensuring that any relevant sections, e.g., related to consent or notice, are conformant.

## 1.2 ORGANIZATION

This guide is organized around the documents provided in the Framework.

The following policy documents are included:

* Responsible Data Science Policy
* Parent Procedure
  + Adjudicative Template
  + Prescriptive Template
* Concepts and Techniques Inventory

A range of approaches or information systems can be used to implement the policy and procedures. Sample process artifacts that can guide implementation or be used by any organization with Office or Google Docs are included:

* Data Science Proposal Form
* Data Science Proposal Review Form
* Data Science Review Log
* Data Science Release Form

## 1.3 UPDATES AND ASSISTANCE

Updates to this framework will be announced via the Responsible Data Science Policy mailing list. Organizations can sign up for this mailing list here:

<https://mailchi.mp/licens.io/responsible-data-science-policy>

Additional resources may be available on the project site here:

<https://licens.io/responsible-data-science-policy/>

Organizations that have comments, questions, or concerns can contact the authors at the following address: [rdsp-support@licens.io](mailto:rdsp-support@licens.io)

# 2. Policy Documents

All documents are provided in standard Office (XML) formats, compatible with Microsoft Office, Google Docs/Sheets, Apache OpenOffice, or LibreOffice.

While organizations should carefully review all documents in their entirety, there are specific noun phrases, verb phrases, sentences, or lists marked for completion in the documents. These sections are marked with single square brackets, e.g., [TO BE COMPLETED].

In other cases, documents include explicitly optional phrases or examples. These optional or example items are marked with double brackets, e.g., [[THIS IS SOMETHING YOU MAY CHOOSE NOT TO DO.]].

## 2.1 RESPONSIBLE DATA SCIENCE POLICY

The Responsible Data Science Policy document is the controlling document for the framework. It defines the guiding principles and purpose and the controlling scope, roles, responsibilities, and procedures.

Organizations should begin by reviewing the four principles listed under the Principles section. These principles have been used to guide the drafting of the entire framework. As such, if the organization desires to remove or modify a principle, they should consider how such a change would affect the rest of the policy, such as the Purpose section immediately thereafter. In particular, while the fourth principle is included to inspire a better world, some organizations may consider its inclusion unhelpful or unnecessary.

Organizations should next review the policy’s Scope definition. Some organizations may have specific business units or organizational structures that can be referenced to provide clarity. Other organizations may classify personnel differently, such as by excluding 1099 individual contractors or by linking the definition to another policy.

Organizations should proceed to consider the roles and responsibilities in the policy. In particular, the policy defines a Policy Owner and Policy Exception Approver. As drafted, both roles may be a single individual or a group (e.g., a committee). However, organizations should consider whether they may desire more or less separation of roles and responsibilities; for example, by combining the Policy Owner and Policy Exception Approver into one role, or by separating responsibility for policy design from policy implementation.

Organizations next need to consider their overall approach to data science review. This framework is designed to cover a wide range of organizations; it is also designed to be able to adapt over time as the activities and risk preferences of an organization change. In particular, the framework is designed to flexibly allow organizations to choose how much delegation they want.

For example, some organizations may prefer strong centralization of risk assessment and decision-making; these organizations will likely rely on an adjudicative process that puts the power in the hands of a committee or board. Other organizations may have well-known or lower-risk processes that recur frequently; as a result, they may select a prescriptive approach that delegates much of the effort downstream to the project personnel performing the work. The parent-child procedure structure even allows for organizations to conditionally direct projects to one or more approaches based on use case, providing maximum flexibility as organizations grow or optimize risk management and productivity.

Once organizations have made an initial choice regarding adjudicative and prescriptive approaches, they should ensure that the Procedures section of the policy reflects their choice. For example, organizations that intend to rely only on one approach may remove sections of the policy that are not applicable to that approach.

Next, organizations should decide which regulations and standards are explicitly referenced or compatible with this policy. This consideration will depend primarily on the geography and industry of the organization, and the organization’s existing compliance efforts related to such regulation. For example, organizations that are subject to GDPR or CCPA should review the list under the Regulations and Standards section and select the appropriate language under the Transparency and Data Use subsections. Additionally, organizations may decide to direct personnel at a Policy Knowledge Base, which might be a wiki or other source of documentation for related compliance programs.

The policy also defines a number of intervals or frequencies related to policy review and covered asset or project review. For example, under the Responsibility section, the policy requires a choice regarding a Policy Review Interval. Similarly, under the Compliance section, the policy requires a choice regarding how frequently each in-scope asset or project is reviewed - the Project Review Interval.

Organizations should also decide if they want to promote these standards through their supply chain. This choice, which may depend on the company’s size, negotiating leverage, industry, and preferences, is implemented in the Supply Chain section of the policy, where a specific Contracting or Procurement Owner role is specified.

Finally, the organization should ensure that the policy “has teeth” in the Violation section by incorporating any relevant enforcement approaches, such as a disciplinary policy.

## 2.2 PARENT PROCEDURE TEMPLATE

The Parent Procedure template is designed to help direct the personnel at an organization to the specific approach that applies to their data science project. This procedure should serve as a “flow chart” or “decision tree” that allows a user to quickly and easily determine which sub-procedure they should subsequently follow.

As drafted, the template provides for three “outcomes” in the flow chart or decision tree:

* Not Permitted
* a Prescriptive Sub-Procedure
* an Adjudicative Sub-Procedure

Organizations can alter these outcomes by removing or adding options. For example, an organization that centralizes the process around a risk committee may remove references to the prescriptive sub-procedure from the document.

Example procedure implementations are provided in both table and diagram formats under the Procedures section. Organizations are strongly encouraged to provide a visual representation of the conditions when possible.

Organizations should also ensure that all contact information and document locations are completed and correct in the document, such as in the Guidance section.

## 2.3 ADJUDICATIVE PROCEDURE TEMPLATE

This document provides a sample template for adjudicative sub-procedures. Most organizations will implement at least one adjudicative sub-procedure; in fact, many organizations will adopt this policy framework by implementing just a single adjudicative body (e.g., a single controlling Data Science Risk Committee).

Once an organization decides what type of adjudicatory body they would like to implement and what the body’s scope is, they should ensure that the document is aligned. For example, an organization that selects an external review board for projects involving minors should consider what changes might be necessary for this approach. In general, most organizations will designate individual reviewers or internal review committees; some organizations may make existing committees responsible, such as a Privacy Committee or Ethics and Compliance Committee.

The adjudicative sub-procedure template is divided into two stages: the Proposal stage and the Approved Project stage. During the Proposal stage, personnel first complete a Data Science Proposal Form that is designed to provide the adjudicator with necessary information for review. A sample form is provided as a spreadsheet in XLSX format in the framework package, but the organization should review this form and adapt it as required. As part of this process, a Liaison is designated for a project to provide a clear point of contact.

The Liaison submits the Proposal Form for the Project Reviewer to review. The submission may simply occur by emailing an attachment, as currently drafted in the template, or could be handled through another information system; either way, the organization should confirm that the process and contact information are complete and accurate under the Project Review section.

The Project Reviewer may be the entire adjudicatory body, e.g., committee, or it may be a single named point of contact for the body. Organizations may select to implement an internal SLA for this review process, as optionally specified under the Project Review section.

As drafted, the Project Reviewer contemporaneously documents their reasoning and responses in a Data Science Review Form. A sample form is provided as a Word document in DOCX format in the framework package, but the organization should review this form and adapt it as required. This form is not intended to be shared outside of the adjudicatory body, though portions of it may be communicated back to the Liaison or Project team if appropriate.

As drafted, outcomes for review include:

* Project approved
* Project conditionally approved - resubmission required
* Project conditionally approved - resubmission not required
* Project denied

Organizations should ensure that the outcomes and related workflows described in the sub-procedure are appropriate for their business and use case as detailed under the Project Review and Conditionally Approved Proposals sections.

The framework is designed around the idea that data science projects typically go through a research and development phase prior to distribution or release. For projects that are one-time or ad-hoc analyses, the “release” may simply be publishing a PDF on a site or distributing a spreadsheet over email. For projects that involve data that may be updated on a recurring basis or models that might become embedded in software applications, “release” may be more aligned with traditional SDLC concepts. Organizations should carefully consider how to define “distribution or release” in the context of their overall business or specific use case.

Once the “research and development” phase of a project is complete and a project is ready to be “distributed or released,” the template adjudicative process requires a second review by the adjudicator. This Approved Projects section mirrors the process from the above Proposal stage. The primary difference is that the Data Science Review Form used in the Proposal stage may not be appropriate for the Release stage. Organizations should carefully consider their business and use case to decide if the review form should be differentiated across stages.

Finally, organizations should ensure that any other relevant policies or guidance are linked to this sub-procedure. For example, the Related Policies and Procedures and Guidance sections allow organizations to list related requirements such as data handling policies or internal compliance knowledge base or wikis.

## 2.4 PRESCRIPTIVE PROCEDURE TEMPLATE

This document provides a sample template for prescriptive sub-procedures. Most organizations will not initially adopt this policy framework with a prescriptive sub-procedure; instead, their adjudicatory body will notice trends and patterns among projects over time. As recurring use cases emerge and clear guidelines are established after research and experience, these guidelines may be formalized into a set of requirements around data, models, and testing - a prescriptive sub-procedure.

For example, an organization with a Data Science Risk Committee may notice that the company is regularly using direct mail to market to potential customers. After approving and subsequently reviewing multiple direct mail underwriting projects, they may memorialize their guidance in the form of a prescriptive approach; going forward, the committee no longer needs to review each individual direct mail campaign so long as the data science team follows the approach defined.

Organizations should begin by clearly describing the applicable scope for a prescriptive sub-procedure in the Scope section. The conditions and requirements should be drafted with the intended audience in mind, so where possible, regulatory requirements should be presented in plain language or clarified for the reader.

Additional requirements or assumptions related to a project’s dataset(s) or pre-trained models can be specified in the Dataset Requirements or Model Requirements sections. In the event that no additional requirements or assumptions are relevant, then these sections can be removed from the document.

Next, the organization should ensure that a Sub-Procedure Owner is specified and that their contact information is stated. Like other roles, this may be a named individual, a monitored mailing list, or a specific committee or other adjudicatory body.

As with adjudicative sub-procedures, prescriptive sub-procedures are divided into two stages: the Research and Development stage and the Release stage. The Research and Development stage begins with the data science personnel noticing the Sub-Procedure Owner of their project. Organizations may consider whether any additional information should be provided at this point.

The framework template organizes the Research and Development stage into the steps below. Organizations should ensure that this structure conforms to their research and development process.

1. Pseudo-Anonymization
   1. Attribute Transformations: This section defines pseudo-anonymization transformations that apply to data subject attributes; in most cases, these are “field” or “column” transformations that do not otherwise alter the data subject records.
   2. Dataset Transformations: This section defines pseudo-anonymization transformations that apply to the dataset as a whole, altering the nature or relationship of records to data subjects. For example, group-by aggregations result in a change in the unit of each record.
   3. Acceptance Criteria: This section defines acceptance criteria for pseudo-anonymization transformations above. Organizations should consider the nature of their business and use case. Ideally, strong differential privacy constraints are used, such as a minimum l-diversity and k-anonymity.
2. Modeling
   1. Systems: This section lists allowed, required, or prohibited systems for managing the modeling process, such as source control systems or artifact repositories.
   2. Cleaning: This section defines allowed, required, or prohibited data cleaning steps.
   3. Augmentation and Synthetic Data Generation: This section defines allowed, required, or prohibited augmentation or synthetic data generation steps.
   4. Feature Engineering: This section defines allowed, required, or prohibited feature engineering steps.
   5. Model Types: This section defines allowed, required, or prohibited modeling techniques or parameters. For example, organizations may only permit models that exhibit certain levels of explainability or determinism.
   6. Training and Testing: This section defines allowed, required, or prohibited training and testing approaches or constraints. For example, organizations may require certain test sample sizes or stratification approaches to address potential bias.
   7. Cross-Validation and Hyperparameter Optimization: This section defines allowed, required, or prohibited cross-validation and hyperparameter optimization techniques and constraints. For example, organizations may require a certain type of k-fold cross-validation or optimization target.

Once the Research and Development stage is complete, the sub-procedure next considers the Release stage. As discussed above, organizations should carefully consider how to define “distribution or release” in the context of their overall business or specific use case.

The framework template organizes the Release stage into the steps below. Organizations should ensure that this structure conforms to their desired release process. For example, some organizations may request additional notice or review in some circumstances prior to distribution or release.

1. Fairness Testing: This section defines requirements for fairness testing for a model, e.g., in situations where disparate impact analysis is required like credit or employment. In some circumstances, such legal or ethical considerations may not be required and this section may be left empty or removed.
2. Productionization Testing: This section defines requirements for testing of the “production” data science systems. For example, in many contexts, a software development or product team may make modifications or translate a model from one language to another for release. As a general rule, organizations should take steps to ensure that such modifications or language “ports” do not alter the data or model results.
3. Online or Periodic Retraining: This section defines requirements for models that are distributed or released as online or periodically-retrained models. Such models may produce different results for the same input over time as new training data automatically triggers refitting of the model. Therefore, additional steps may need to be taken to allow for point-in-time audit or historical review.

Finally, organizations should ensure that any other relevant policies or guidance are linked to this sub-procedure. For example, the Related Policies and Procedures and Guidance sections allow organizations to list related requirements such as data handling policies or internal compliance knowledge base or wikis.

## 2.5 CONCEPTS AND TECHNIQUES INVENTORY

This document is intended to provide a baseline inventory of concepts and techniques that are used in the framework template or are commonly used in practice. This inventory is neither exhaustive nor technically precise in nature, but should serve as an adequate reference for initial adoption. Over time, organizations should monitor industry trends or emerging risks, such as hash algorithm safety, to ensure that the document is up-to-date.

Organizations should also consider whether undesirable techniques should be completely removed from the inventory or whether they should be kept for reference when counterparties or systems rely on such otherwise-prohibited techniques.

Furthermore, organizations are encouraged to map these concepts and techniques to specific open source or internal software implementations. For example, many transformations or models have implementations in languages or libraries like SQL-99, R, Apache Spark, or the pandas or scipy libraries in Python. This is also true for pseudo-anonymization, differential privacy, and homomorphic encryption, where open source libraries like OpenDP, Google’s differential-privacy, Microsoft’s SEAL can provide reference implementations.

In the extreme case, organizations can modify their prescriptive and adjudicative sub-procedures to directly link procedure steps to specific software implementations, data sources, or APIs.

# 3. Sample Artifacts

The policy framework includes five sample artifacts related to the proposal, review, and release steps of the policy sub-procedures. These are intended to be demonstrative and organizations should consider how to modify or extend them for their business or use cases.

All artifacts are provided in standard Office (XML) formats, compatible with Microsoft Office, Google Docs/Sheets, Apache OpenOffice, or LibreOffice. However, organizations are encouraged to consider whether these forms and logs are better implemented in another system, such as an existing governance or compliance system.

## 3.1 POLICY EXCEPTION LOG

This sample form demonstrates how an organization might track exceptions to the policy and procedures.

## 3.2 DATA SCIENCE PROPOSAL FORM

This sample form demonstrates how an organization might collect information from data science personnel prior to initial review by an adjudicative body like a committee or review board. Organizations should consider whether the terminology in this form conforms to their use. Organizations should also consider whether additional questions should be asked or whether some fields should be removed, especially if the proposal is used for specific use cases or regulatory requirements.

Organizations should consider the appropriate user experience for this form during implementation. The provided template has been formatted to allow for the widest compatibility between software systems. For example, the checkboxes are not interactive. Organizations may opt to replace the boxes with interactive checkboxes in Excel, or to implement the form in a web application or other compliance system.

## 3.3 DATA SCIENCE REVIEW FORM

This sample form demonstrates how an organization might document the reasoning and conclusions of an adjudicative body like a committee or review board. Organizations should consider whether the terminology in this form conforms to their use. Organizations should also consider whether additional questions should be asked or whether some fields should be removed, especially if the review is related to a specific use case or regulatory requirements. In many cases, organizations may need to create separate review forms for different use cases.

## 3.4 DATA SCIENCE REVIEW LOG

This sample form demonstrates how an organization might track the inventory and status of all projects and related reviews that have been handled under this policy. Organizations should consider whether the terminology in this form conforms to their use. Organizations should also consider whether additional fields should be added to this inventory.

## 3.5 DATA SCIENCE RELEASE FORM

This sample form demonstrates how an organization might collect information from data science personnel prior to final review by an adjudicative body like a committee or review board. Organizations should consider whether the terminology in this form conforms to their use. Organizations should also consider whether additional questions should be asked or whether some fields should be removed, especially if the proposal is used for specific use cases or regulatory requirements.

Organizations should consider the appropriate user experience for this form during implementation. The provided template has been formatted to allow for the widest compatibility between software systems. For example, the checkboxes are not interactive. Organizations may opt to replace the boxes with interactive checkboxes in Excel, or to implement the form in a web application or other compliance system.