



Department for
Science, Innovation
& Technology

The Model for Responsible Innovation

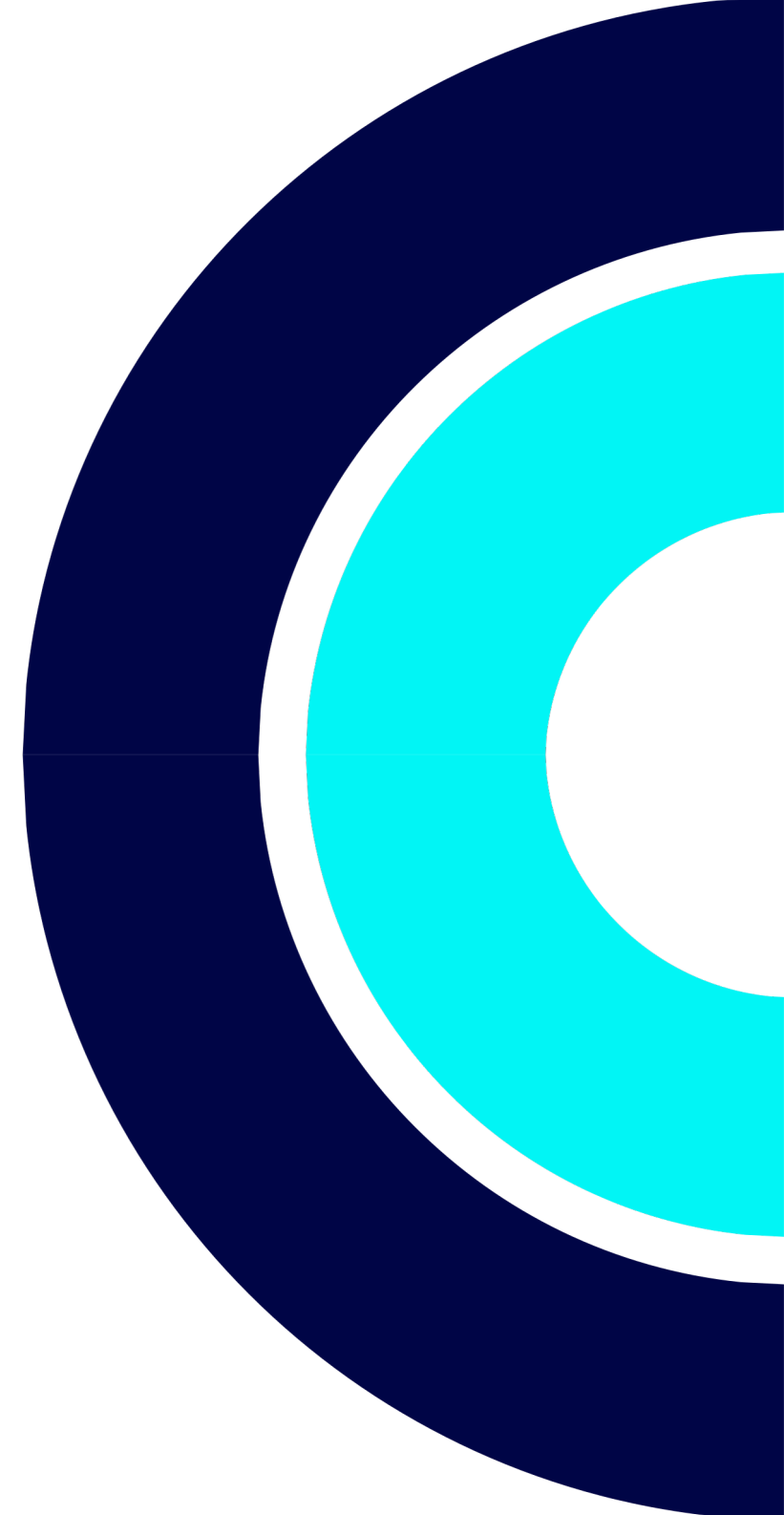
A practical tool for trustworthy
adoption of AI in the public sector

Autumn 2024



Contents

01 Executive Summary	3
02 Introducing the Model	5
03 The Model in detail	10
Trustworthiness	11
The Fundamentals	12
The Conditions	14
04 Our Offer: The Model Workshop	17





01

Executive Summary



The Model for Responsible Innovation

The Model for Responsible Innovation is a practical tool created by DSIT's Responsible Technology Adoption Unit (RTA) to help teams across the public sector and beyond to innovate responsibly with data and AI.

The RTA uses the model to run **red-teaming workshops** with teams developing data-driven technology.

These workshops are run by RTA experts and free for public sector teams. They provide rapid, hands-on support to help teams identify the biggest threats to trustworthiness in their current use cases, and tailored recommendations for addressing them.

The Model for Responsible Innovation can help you if:

- You are a **public sector team** delivering a project which contains some element of data-driven technology or AI
- You are a **private sector team** building an AI or data-driven tool which you plan to use for a public sector purpose, or has a significant societal footprint

Please reach out to us at rtau@dsit.gov.uk to register your interest in a free session.





02

Introducing The Model

What is the Model?



The Model for Responsible Innovation is a practical tool created by DSIT's Responsible Technology Adoption Unit (RTA) to help teams across the public sector and beyond to innovate responsibly with data and AI.

The Model does two things:

- **It sets out a vision for what responsible innovation in AI looks like,** and the component Fundamentals and Conditions required to build trustworthy AI
- **It operates as a practical tool** that public sector teams can use to rapidly identify the potential risks associated with the development and deployment of AI, and understand how to mitigate them.

The Responsible Tech Adoption Unit uses the model to run **red-teaming workshops** with teams seeking to innovate with data-driven tech. These collaborative sessions map data and AI projects against the model to rapidly identify where risks might arise, and prioritise actions to ensure their approach is trustworthy.

We are now making the Model publicly available to enable more teams to take advantage of this approach. Using the Model to build trustworthiness into AI tools across the public sector will help the UK to innovate with data-driven technologies whilst addressing the risks and building trust.



Why do we need the Model?

AI is already providing significant and society-wide benefits, from medical advances to mitigating climate change. The UK needs to capitalise on the huge benefits of these technologies to deliver economic growth and improved public services.

However, we should also not overlook the new risks that may arise from the use of AI tools. These risks threaten to undermine public trust in AI, and hold us back from seizing the opportunities the technology can offer. **By building trust, we can accelerate the adoption of AI across the UK to maximise the benefits whilst addressing the risks.**

The RTA has built our Model for Responsible Innovation over several years of research and testing, as a framework that enables a structured conversation about responsible innovation in data and AI, and can be

practically used by public sector teams to identify risks and mitigations.

The Model builds on existing frameworks and principles for ethical AI, such as the [OECD principles for trustworthy AI](#). It is designed to align with the UK's existing domain-specific guidance for responsible innovation, such as the [Data Ethics Framework](#).

It breaks down responsible innovation into **three main components: the Fundamentals, the Conditions, and the central goal of Trustworthiness**. This guide provides an introduction to each, explaining why they are relevant and how they work together.

How can the Model help you?

The Model for Responsible Innovation can help you if:

- You are a **public sector team** delivering a project which contains some element of data-driven technology or AI
- You are a **private sector team** building an AI or data-driven tool which you plan to use for a public sector purpose, or has a significant societal footprint

The Model can be used at **any point in the development and deployment lifecycle**, but is most useful:

- **At the beginning of a project**, when deciding whether data-driven technology could help solve a policy or delivery problem
- When you have chosen a solution and you are **beginning the development process**
- When you have built a tool and are **deciding how to deploy it.**

The Model can help you:

- By providing a **framework of issues you should consider** during development
- By helping you think about how to **set up ways of working** and governance across the project lifecycle
- Through our **Red-Teaming workshops**, by directly testing your project against the Model to assess whether you are building your tool in a responsible way.

How did we build the Model?

The Model is the product of several years of design, iteration and testing by the RTA.

We designed the Fundamentals by carrying out a comprehensive mapping of public sector data ethics frameworks and principles, such as the [OECD principles for trustworthy AI](#). We identified the unifying components and synthesised them into the core eight Fundamentals. This was further iterated by applying the Model to past RTA projects, such as our work on the [MOD's AI Ethics Principles](#).

We designed the remaining elements of the Model by identifying the core underlying conditions that public sector teams needed to have in place to build responsible AI, and matching them to the interventions RTA have carried out to directly assist teams across Government.

Since its initial design, the Model has gone through several design changes, and substantial testing against projects across the full range of public sector use cases. We have used the Model with teams from policing to social care, from education to transport.

We are confident that this version of the Model provides the most efficient and accessible way to understand and build responsible innovation in data projects across the Government space. **We will continue to test and update the Model over time so that it remains as accessible as possible and in line with latest best practice.**



03

The Model in Detail

Trustworthiness



At the centre of the Model is **trustworthiness**. The objective of responsible innovation is building **justified trust** in the AI and data tools we develop and use.

Justified trust comes through designing and deploying systems in a way that **builds and deserves the trust** of stakeholders to use them. Without confidence that a system has been built and deployed responsibly, the public are unlikely to support and use it, and its full potential may not be realised.

Therefore, the different elements of the Model are carefully designed to help build towards this goal. By following the Fundamentals and Conditions, teams can build their systems so that they earn the justified trust of those using them.

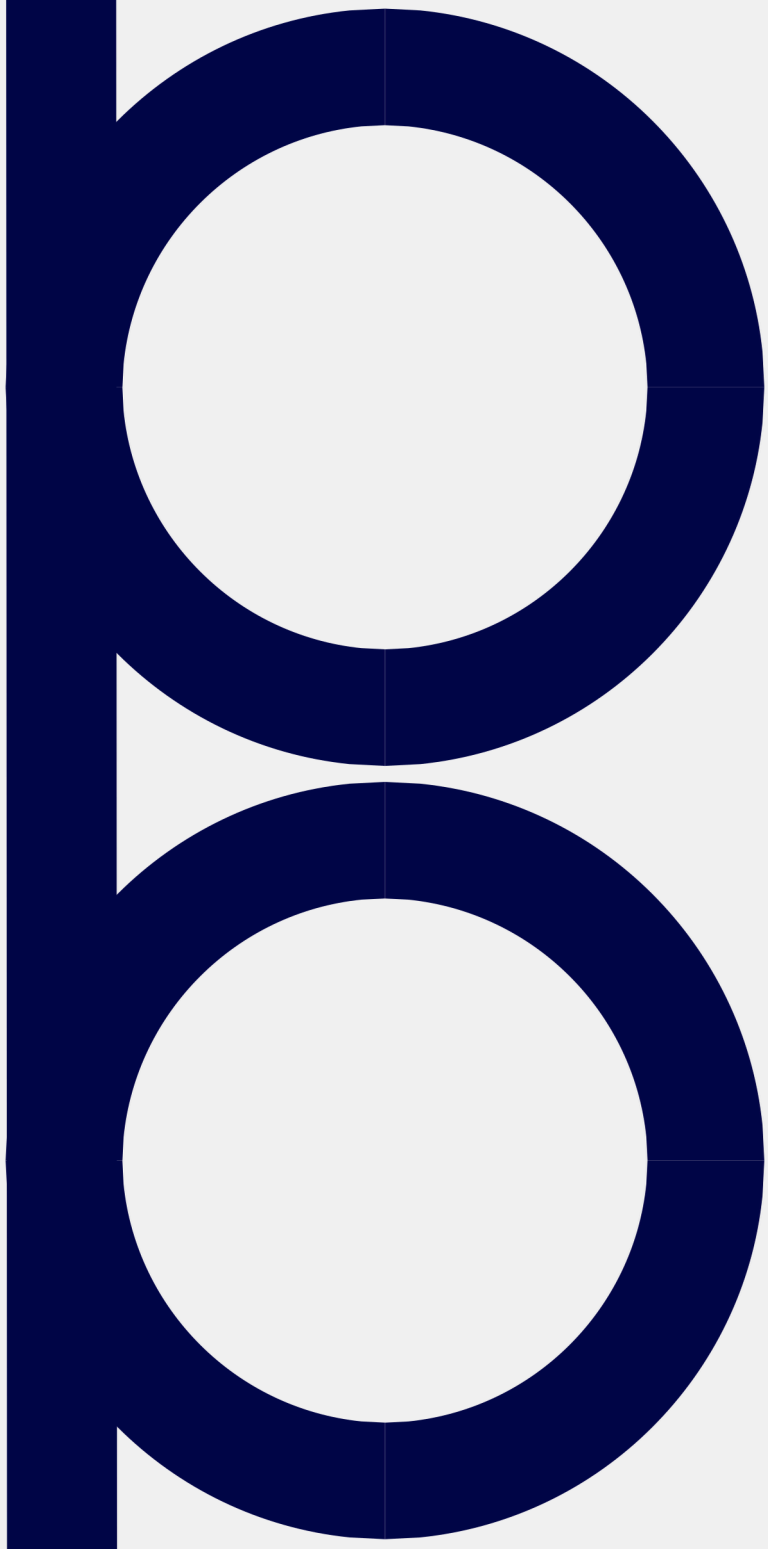


The Fundamentals

To build this trustworthiness, the Model outlines eight Fundamentals that teams should work towards when developing and implementing their systems. These operate in part as principles to follow, but also as lenses to consider the broad range of ethical risks that might emerge when carrying out a complex data-driven project. They include:

- **Transparency** – ensuring systems are open to scrutiny, with meaningful information provided to relevant individuals across their lifecycle.
- **Accountability** – ensuring systems have effective governance and oversight mechanisms, with clear lines of appropriate responsibility across their lifecycle.
- **Human-centred Value** - ensuring systems have a clear purpose and benefit to individuals, and are designed with humans in mind.
- **Fairness** - ensuring systems are designed and deployed against an appropriate definition of fairness, and monitored for fair use and outcomes.
- **Privacy** - ensuring systems are privacy-preserving, and the rights of individuals around their personal data are respected
- **Safety** - ensuring systems behave reliably as intended, and their use does not inflict undue physical or mental harms.
- **Security** - ensuring systems are measurably secure and resistant to being compromised by unauthorised parties.
- **Societal Wellbeing** - ensuring systems support beneficial outcomes for societies and the planet.





Whilst all the Fundamentals should be present in any use of AI, they do not necessarily all need to be maximised, as there will often be occasions where emphasising one Fundamental will bring trade-offs with another. For example, some projects could maximise security, at the cost of increased risks that the system is less explainable, transparent and accountable.

In our red-teaming workshop approach, it is by assessing a project against each Fundamental that we can identify where the greatest threats to trustworthiness lie, whether that's from poor accountability practices or the risk of unfair outcomes. We can then help teams consider how best to manage these trade offs.

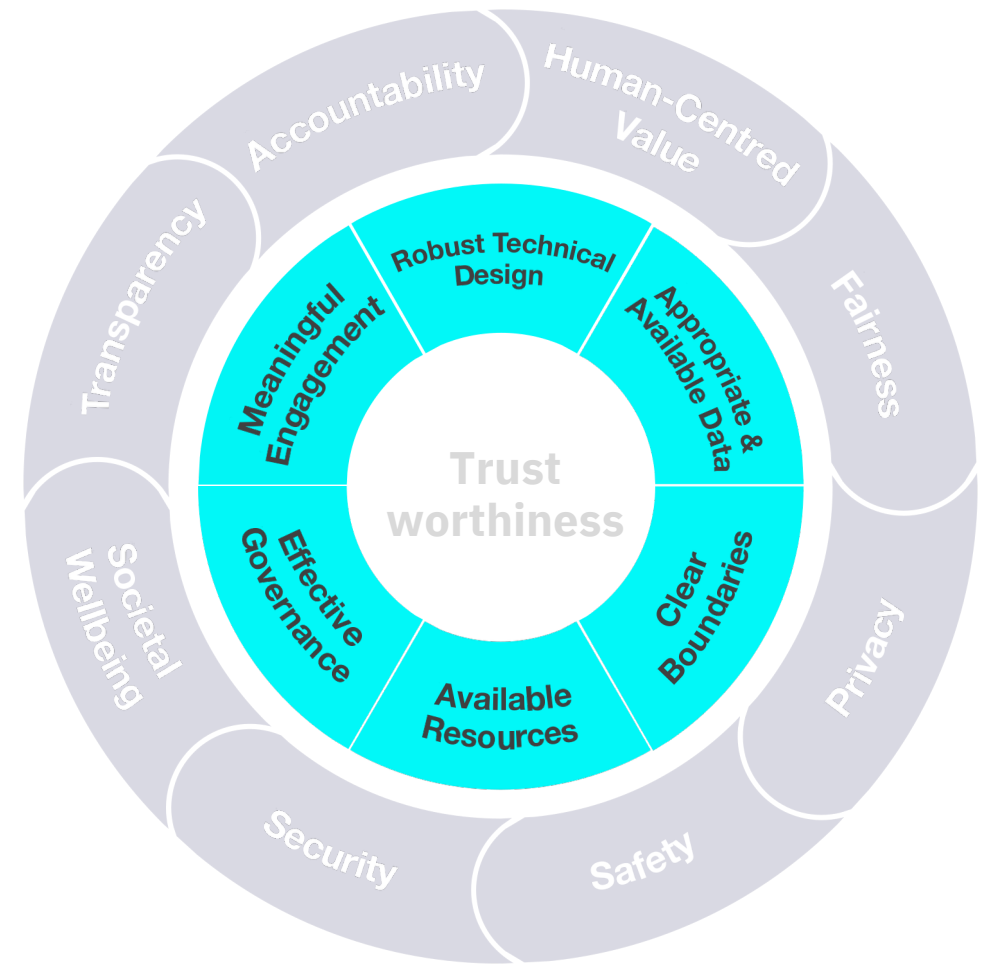
The Fundamentals are underpinned by requirements in UK law, including UK GDPR, the Equality Act 2010 and other legislation. These set a baseline requirement that all development teams need to comply with when delivering AI and data projects. Red-teaming workshops using the model both give an opportunity to identify areas of law that are relevant, but also to explore good practice beyond the legal minimum, and where trade-offs exist between Fundamentals.


The Conditions



Underlying the Fundamentals are the Conditions. These are the technical, organisational and environmental factors that must be satisfied in order for the Fundamentals to be met. Located on the inner ring of the Model, they are:

- **Meaningful Engagement** - engaging effectively with experts, stakeholders, and the general public, using these insights to inform the system in question.
- **Robust Technical Design** – ensuring that the functional (how a program will behave to outside agents) and technical (how that functionality is implemented in code) design of a system is robust.
- **Appropriate & Available Data** - ensuring a system has access to the right data needed to achieve its desired outcomes and effectively monitor performance.
- **Clear Boundaries** - ensuring there are clear boundaries on a system's intended use, and clear understanding of the consequences of exceeding them.
- **Available Resources** – ensuring the resources (technical, legal, financial, etc.) needed to effectively build and use a system are provided.
- **Effective Governance** – ensuring that the right processes and policies are in place to guide the development and operation of a system, and ensure its adherence to the project's goals, standards and regulations, providing recourse where necessary.





These six categories capture different types of measures that teams can take to mitigate the ethical risks present in their projects.

For instance, upholding the Fundamental of Transparency requires meaningful engagement with stakeholders about how an algorithm works, such as filling in an [Algorithmic Transparency Record](#). Whilst ensuring Fairness depends on representative, accurate and up-to-date training data that can mitigate bias, and rigorous testing to ensure that a system operates as intended and is technically robust.

Therefore, the Conditions provide a valuable structure for helping teams respond to the issues highlighted by the Fundamentals.

Underlying Themes

The Model sets out the key Fundamentals and Conditions requires for responsible innovation.

However, there are a number of key themes underpinning the Model which should be considerations for any team developing AI in the public sector. These include:

Legal Compliance

Compliance with the law is a necessary, but not sufficient, element to achieving trust in the use of AI.

Understanding

Public sector organisations must have teams with the right understanding of the technology they are developing, with suitably trained or qualified individuals.

Continuous Evaluation

Reflecting that AI systems are not static and often require an ongoing approach to risk and harms mitigation.

Organisational Culture

Any team developing an AI tool should have a culture which enables and values responsible innovation.



Department for
Science, Innovation
& Technology

04

Our Offer: The Model Workshop

The Model Workshop

What is it?

The Model is designed to be used as part of the RTA's dedicated **ethical red-teaming workshop**, which has been specifically created to guide public sector teams through ethical risks associated with their own projects.

The workshop is run by RTA experts and **free for public sector teams**. It provides hands-on support to help teams identify the biggest threats to trustworthiness in their current use cases, and provides tailored recommendations for addressing them in a short report.

Who is it for?

Any public sector team who has identified a use-case for AI that you want to implement, or have already begun the development process, and want to ensure that it is responsible and robust.

The Model will also be of interest to private sector teams developing a data-driven technology tool for government, or a tool which will have a significant impact on wider society.

RTA has carried out workshops with a range of stakeholders to identify ethical risks in their projects, from central government teams to local councils.

Please reach out to us at rtau@dsit.gov.uk to register your interest in a free session.

What to Expect

The Model red-teaming workshops are a **semi-structured assessment** of the tool you are developing with the RTA's experts. Ahead of the workshop, we'll ask you to complete a **warm-up sheet** with key details and background to the project.

The workshop itself is a **series of discussions based on the Model's Fundamentals**. We will bring prompt questions to provoke discussion and potentially flag risks or issues you hadn't considered, and work together to discuss possible answers.

For example, for **Transparency**, some potential prompt questions are shown here. Some workshops may then follow a similar format for the **Model Conditions**, answering from an organisational level. Sometimes, this will require a separate workshop.

You don't need to be a technical expert or senior leader to participate in the workshops – the workshops aim to facilitate an open discussion that values all perspectives on a project, and provides a rapid and rounded assessment of the ethical implications of any AI tool.

The RTA team will write up the discussion and any key risks identified into a **short actionable report** for the team.

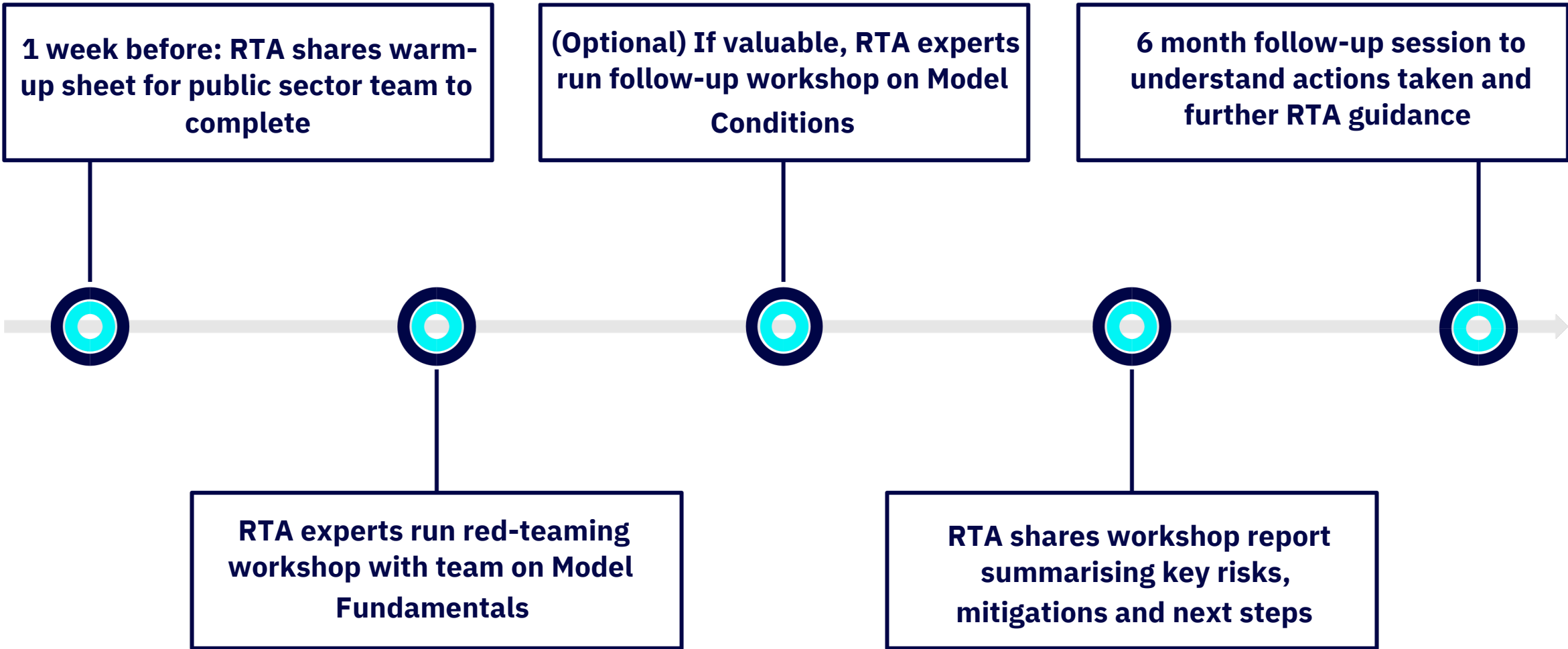
Transparency: systems are open to scrutiny, with meaningful information provided to individuals across their lifecycle.

Will the project be understandable by non-technical users? How will you achieve this?

Will users be aware they are interacting with an AI?

Will you tell users of the system how outcomes were reached? How will you communicate this?

Workshops Timeline



Case Study

Better Outcomes through Linked Data (BOLD)



Better Outcomes through Linked Data (BOLD) is a data-linking programme which aims to improve the connectedness of Government datasets across justice, health and local Government.

The RTA ran a series of red-teaming workshops with teams from BOLD's 4 pilot projects, using the Model.

Following these workshops the RTA drafted an 'ethical risk assessment' summarising the key risks identified across the programme, and recommended actions to help mitigate each.

As a result, the BOLD team developed a new approach to governance and ethical risks, as well as taking steps to improve the documentation of their datasets.

Case Study

DESNZ Project Delivery Chatbot



Department for
Energy Security
& Net Zero

The Department for Energy Security and Net Zero (DESNZ) are developing an AI-enabled chatbot to assist project delivery experts by providing advice, ideas and support based on departmental best practice documents.

The RTA carried out a Model red-teaming session with the DESNZ development team, and identified challenges around how to avoid using sensitive data in the chatbot outputs, and preventing biases from training data.

The workshop has now allowed the DESNZ team to incorporate these risks into their system development.



Department for
Science, Innovation
& Technology

