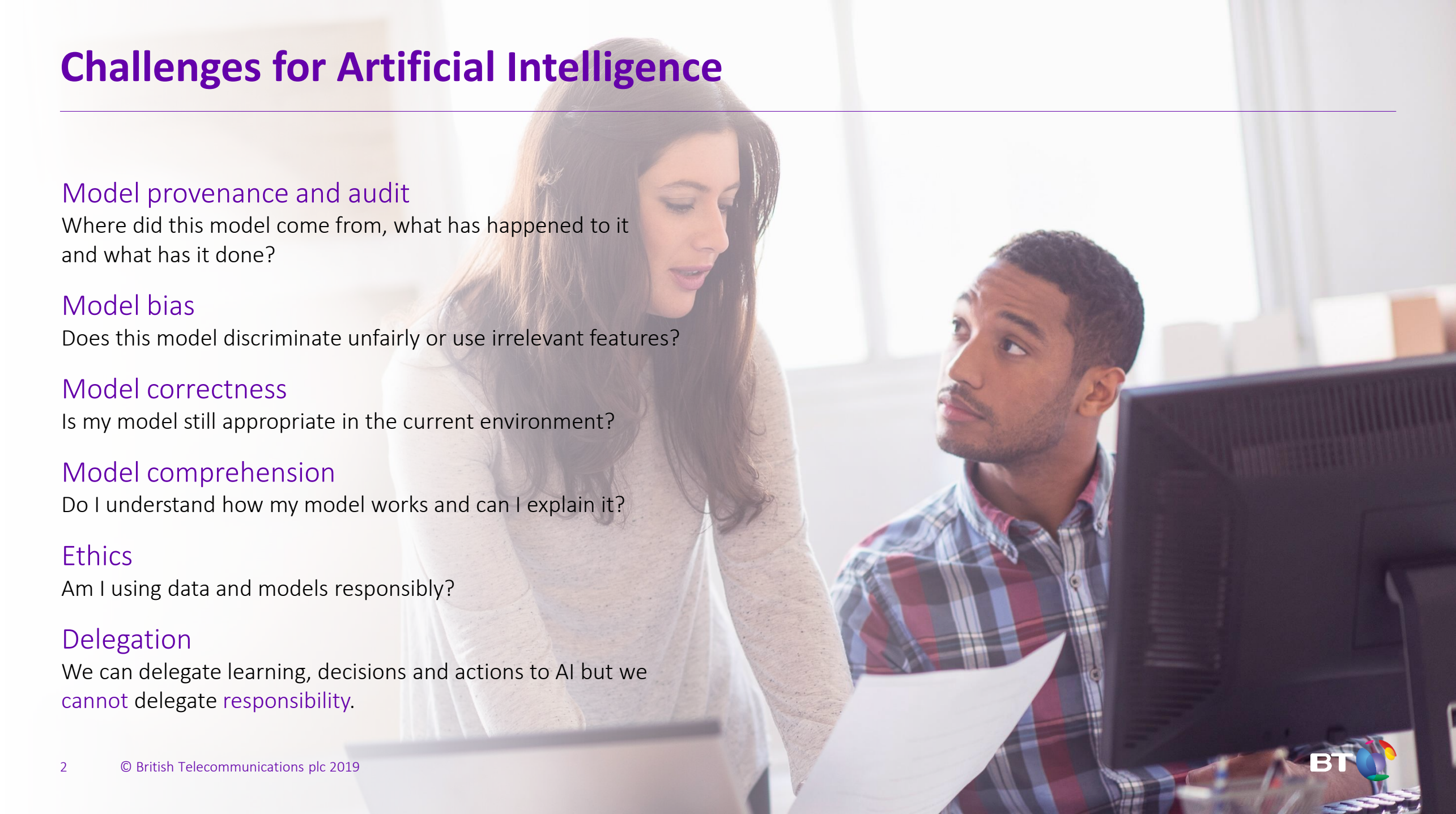
AI Governance

Rob Claxton
Senior Manager in Research, Big Data, Insight and Analytics

Challenges for Artificial Intelligence



Model provenance and audit

Where did this model come from, what has happened to it and what has it done?

Model bias

Does this model discriminate unfairly or use irrelevant features?

Model correctness

Is my model still appropriate in the current environment?

Model comprehension

Do I understand how my model works and can I explain it?

Ethics

Am I using data and models responsibly?

Delegation

We can delegate learning, decisions and actions to AI but we cannot delegate responsibility.

Why is AI different?

Software engineering → Procurement → IT deployment and lifecycle management

+ [• Opaque model • Evolving • Autonomous.] + Composition + Scale + Regulatory / ethical focus

Why is AI different?

Class	Accuracy
Brabancon griffon	1.12
Griffon	1.11
Dog	0.82
Domestic animal	0.75
Canine	0.73

Class	Accuracy
Sporting dog	1.28
Dog	1.24
Domestic animal	1.13
Hunting dog	1.12
Canine	1.09

Procurement → IT deployment and lifecycle management → (back to Procurement)

+ Composition + Scale + Regulatory/ ethical focus

Why is AI different?

Software engineering → Procurement → IT deployment and lifecycle management

+ [• Opaque model • Evolving • Autonomous.] + Composition + Scale + Regulatory / ethical focus

Hypothetical scenario one

An organisation has built an AI model for credit checking and deployed many instances of it across the enterprise.

After 12 months of operation, it becomes known that the training data used to develop the model has been poisoned.

Carefully manipulated input data can force the model to output an incorrect classification.

- How do we find all the affected models?
- Have any of the AI models been attacked?
- Have customers been adversely affected?
- Which ones?
- How do we prevent reoccurrence?

Hypothetical scenario two

A Telco is using AI systems to optimise operation of its 5G networks.

A network outage occurs impacting communications UK-wide, including emergency services.

A malicious exploit of the AI systems is suspected.

Government demands an explanation for the outage and demonstration of appropriate controls for ensuring safe and correct operation of AI for network control.

- How do we identify all model inputs – training data, model weights, embeddings?
- Where are the results of development / test runs for the AI?
- How can we show proper governance for the AI lifecycle including acceptance into service?
- What was the chain-of-custody for the AI throughout its lifecycle – who did what, when?
- How do we explain actions taken by a ‘black-box’ AI system?

Hypothetical scenario three

A data scientist develops a chat bot for internal use.

The chat bot is repurposed by a separate team and deployed on an external website to handle customer enquiries.

The training data was not thoroughly cleaned and it's possible to provoke the chat bot to reveal personally identifying information.

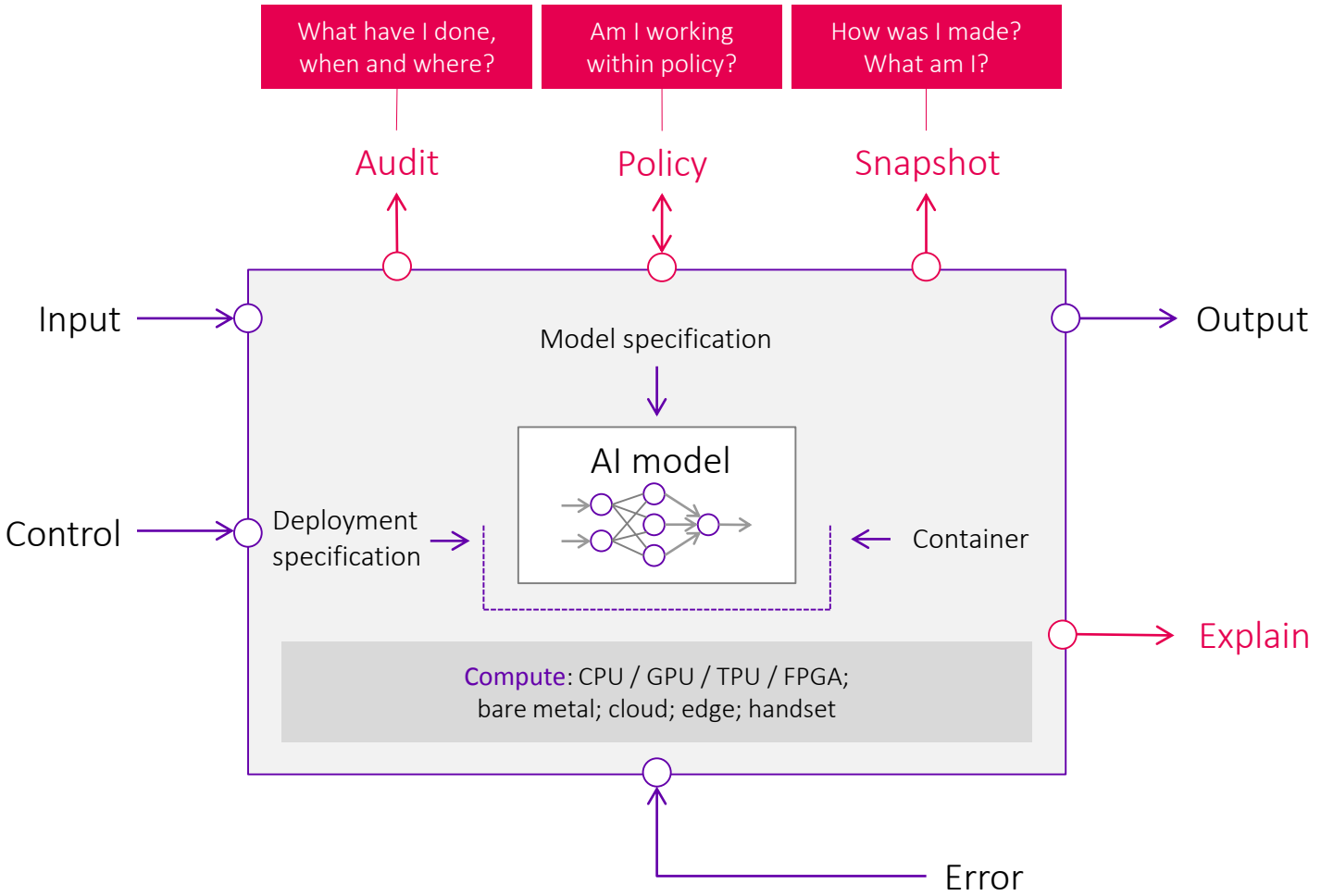
The phone rings – the ICO wants a word…

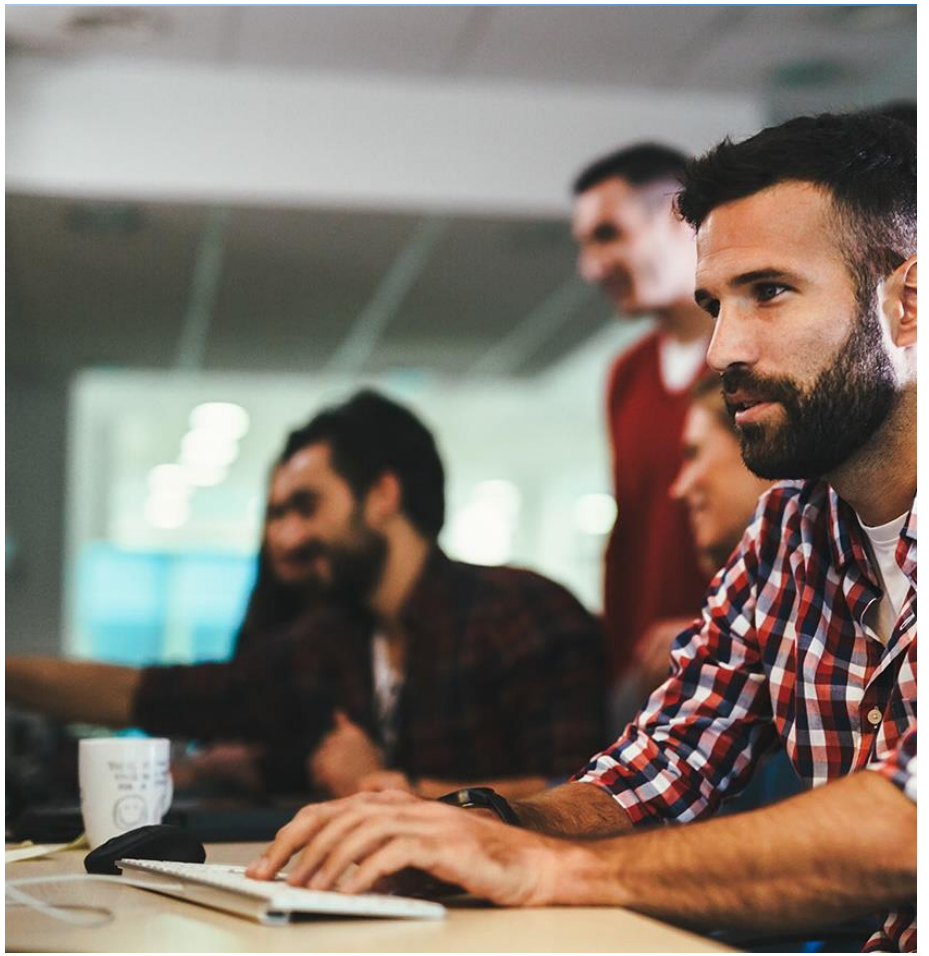
All fingers are pointing at the data scientist who developed the chat bot.

FOR INDICATION PURPOSE ONLY

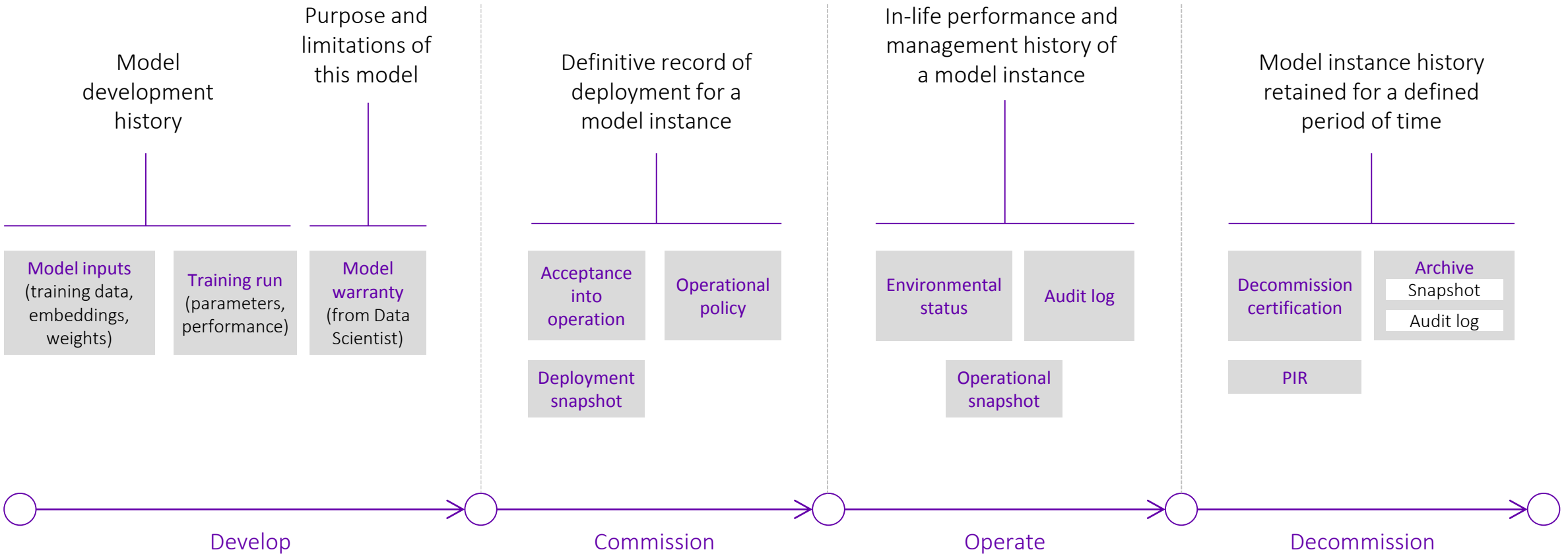
- How can the data scientist demonstrate that the model was not intended for external use?
- How can the limitations on the use of a model be made explicit and preserved for the life of that model?
- How can we test if our models comply with current business policy?

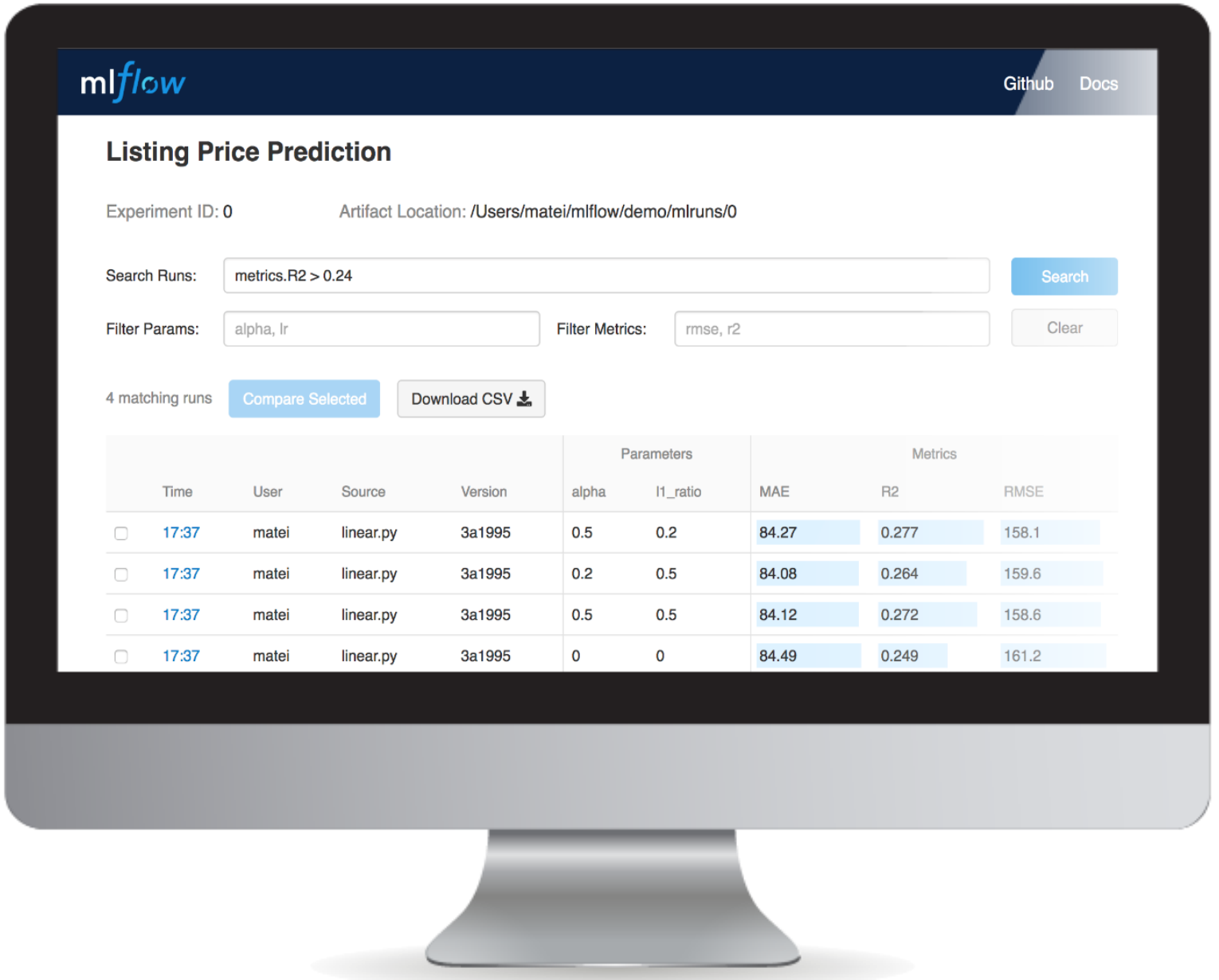
Managing AI

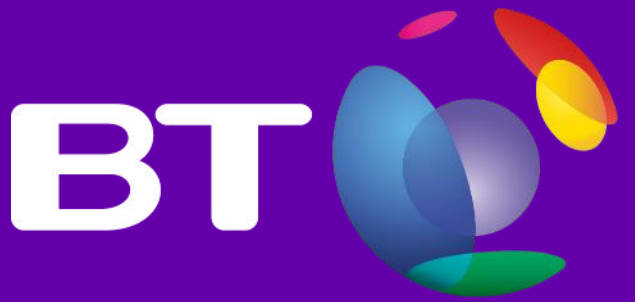




AI lifecycle






BT