The Alan Turing Institute

AI Ethics and Governance

Day 5: Transparency, explainability and the CARE & Act principles

25/11/2022

Dr. David Leslie
Director of Ethics and Responsible Innovation

OVERVIEW

Introduction to Practical Ethics	AI Harms & Values	AI Sustainability	Fairness & Bias Mitigation & Accountability	Transparency, & the CARE & Act principles
01	02	03	04	05

OVERVIEW

Introduction to Practical Ethics

AI Harms & Values

AI Sustainability

Fairness & Bias Mitigation & Accountability

Transparency, & the CARE & Act principles

01 02 03 04 **05**

CONTENTS

Day 4 recap

01 Accountability & Transparency

Q&A

Activity 1: Information gathering and evaluating explanations

Lunch break

02 CARE

02 ACT

Q&A

Final reflections

Recap from Day 4

Recap from Day 4

Fairness in AI

Recap from Day 4

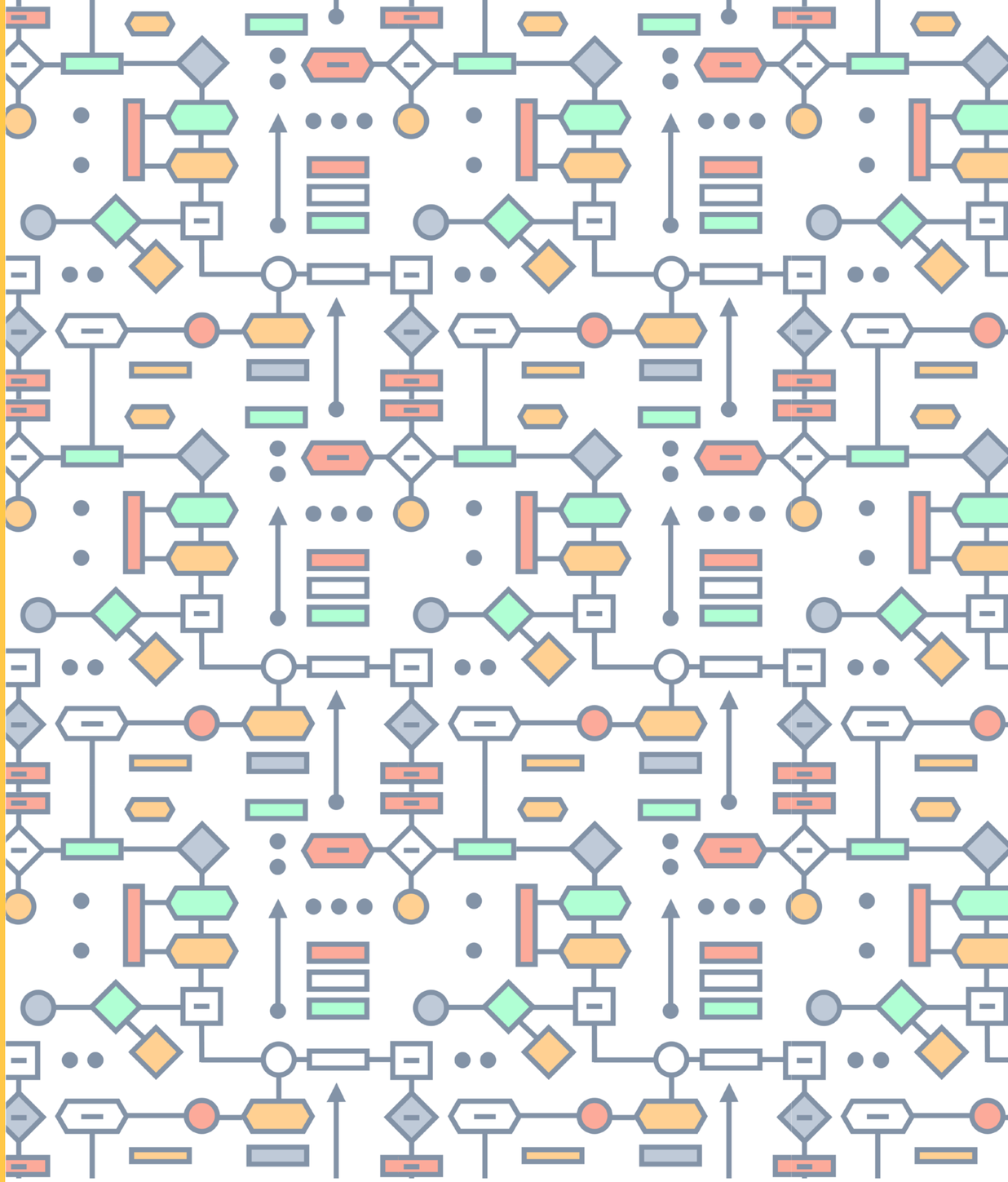
- Fairness in AI
- Bias mitigation

Recap from Day 4

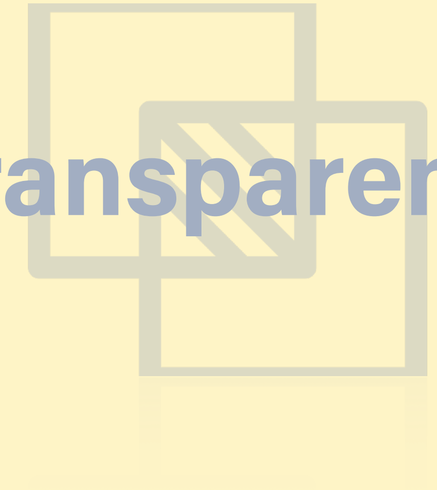
- Fairness in AI
- Bias mitigation
- Accountability

TRANSPARENCY AND EXPLAINABILITY

01



What is transparency in AI?



What is transparency in AI?

1/ the quality an object has when one can see clearly through it

2/ the quality of a situation or process that can be clearly justified and explained because it is open to inspection and free from secrets

What is transparency in AI?

Interpretability & Explainability

Overall Justifiability

Two meanings for transparency in AI

Interpretability & Explainability

- **Interpretability** involves the ability to know exactly **why** and **how** a modern performed the way it did in a certain context. Full access to the rationale behind an output

- Where an AI model is too complex to be understood by human-scale reasoning, Explainable AI methods **(Explainability**) are used to try to '**open the black-box**.'

Two meanings for transparency in AI

Interpretability & Explainability

- **Interpretability** involves the ability to know exactly **why** and **how** a modern performed the way it did in a certain context. Full access to the rationale behind an output
- Where an AI model is too complex to be understood by human-scale reasoning, Explainable AI methods **(Explainability**) are used to try to '**open the black-box**'.

Justifiability

- **Justifiability** involves involves demonstrating that both processes behind and the outcomes of the use of the AI system are ethically justifiable (i.e. sustainable, safe, fair, and driven by responsibly managed data)
- **Justifiability** entails moving beyond the technical clarification of mathematical outputs, demanding a holistic approach that looks at the responsibility of processes and the ethical permissibility of outcomes

How can we explain an AI system to impacted people?

Two kinds of explanations:

1. Process-based explanations

Explanations in terms of the design and implementation practices that lead to an AI-supported outcome

Show that good governance processes and best practices have been followed throughout the design, development, and use of the AI system

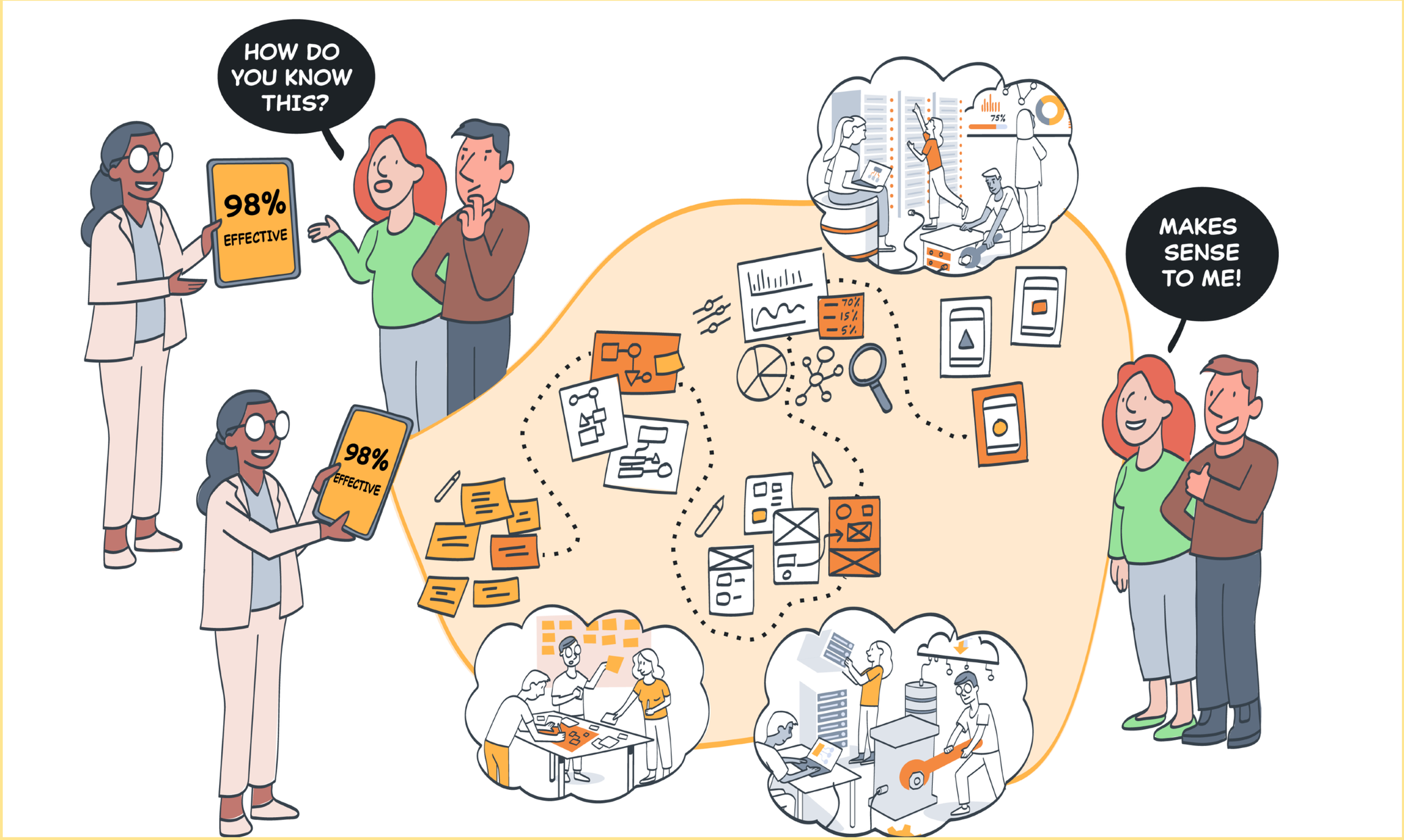
2. Outcome-based explanations

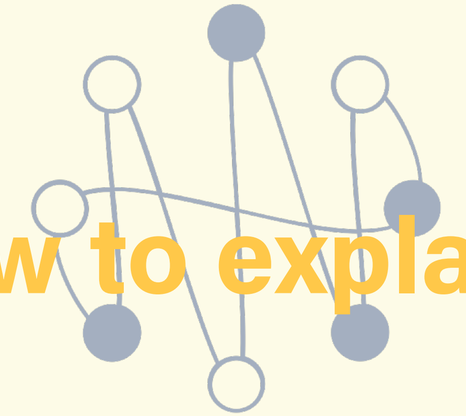
Explanations in terms of the content and justification of the outcome itself

They are about clarifying the results of a specific decision

These outcomes should be communicated in plain, easily understandable, everyday language that is socially meaningful for impacted stakeholders

Building explainable AI systems





How to explain?

Main types of explanations

Rationale explanation

Responsibility explanation

Data explanation

Fairness explanation

Safety and peformance explanation

Impact explanation

Rationale explanation

- It is about the '**why**?' of an AI decision
- Clarifies the **reasons** that led to a decision, delivered in an **accessible** and non-technical way
- Demonstrates how the system got to a result, by showing its **underlying logic**—i.e how it mapped its inputs to its output
- Shows how the system's results apply to the **concrete context and life situation** of the affected individual
- What **purpose** does this explanation serve?
  - Challenging a decision
  - Changing behaviour

Responsibility explanation

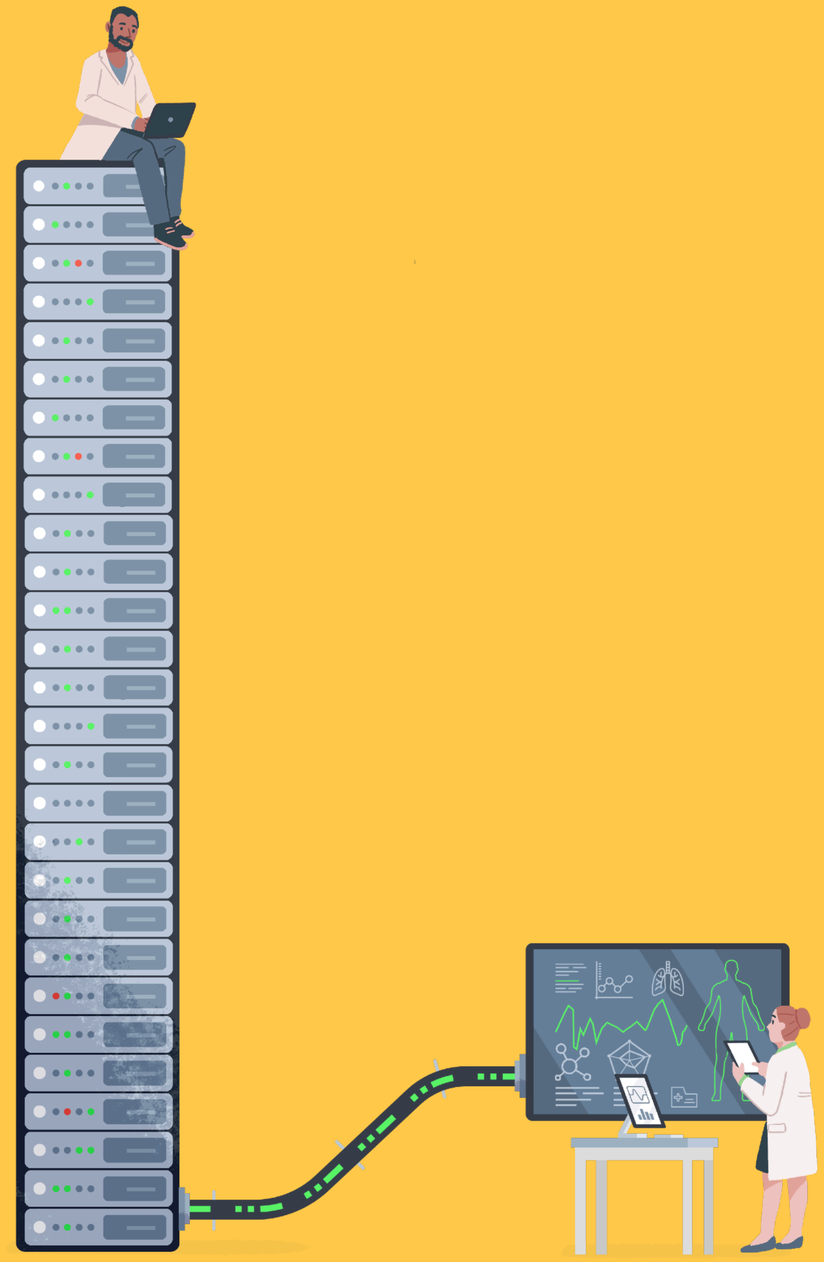
- It is about ‘**who?**’ is involved in the development and management of the AI model
- It tells people who to contact for a **human review** of the decision and how to pursue effective remedy
- What **purpose** does this explanation serve?
  - Challenging a decision
  - Informative purposes

Responsibility explanation

What this type of explanation could include:

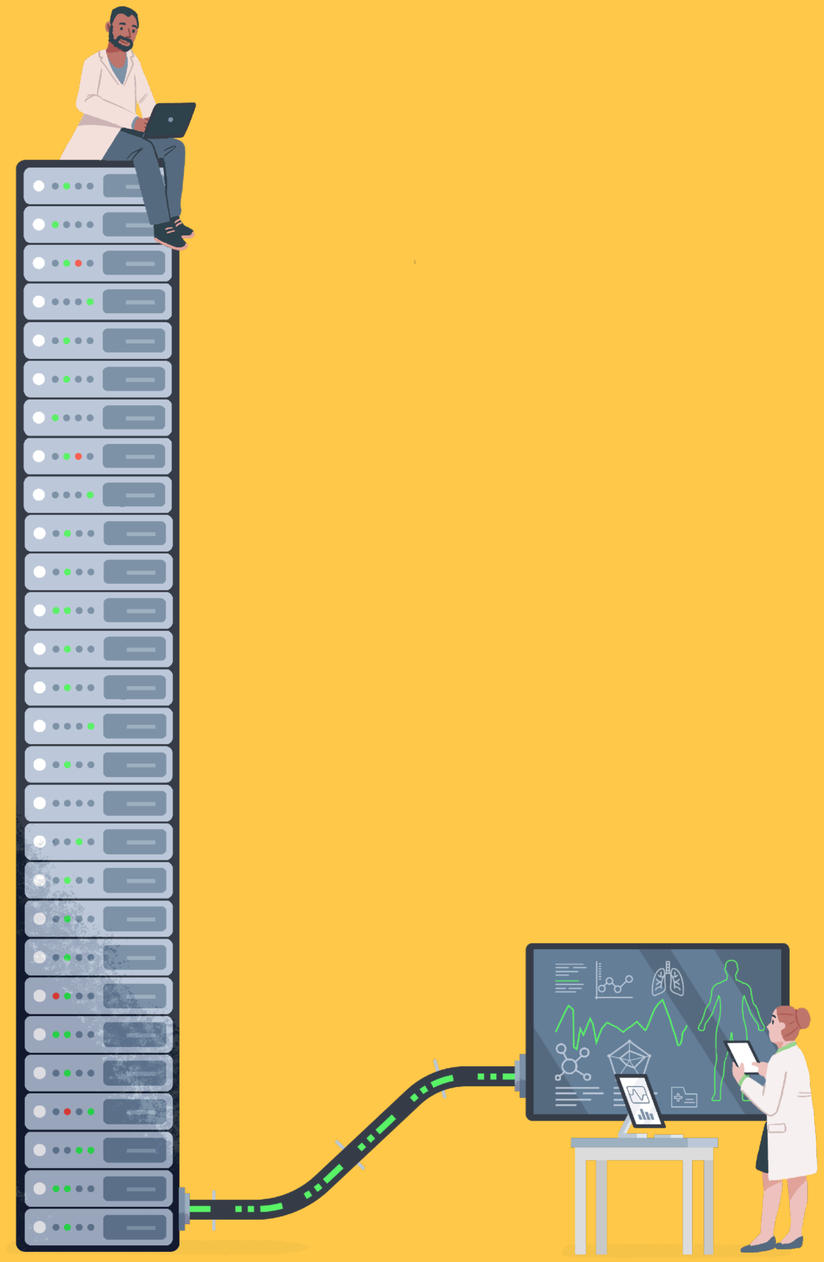
- Information about who is accountable at each stage of the AI system's design, development, and deployment, from project planning and defining outcomes for the system at its initial phases of planning and design, through to providing the explanation to the affected individual at the end.

- Clarification of the mechanisms by which each of these people will be held accountable, as well as how the design and implementation processes of the AI system have been made traceable and auditable.



Data explanation

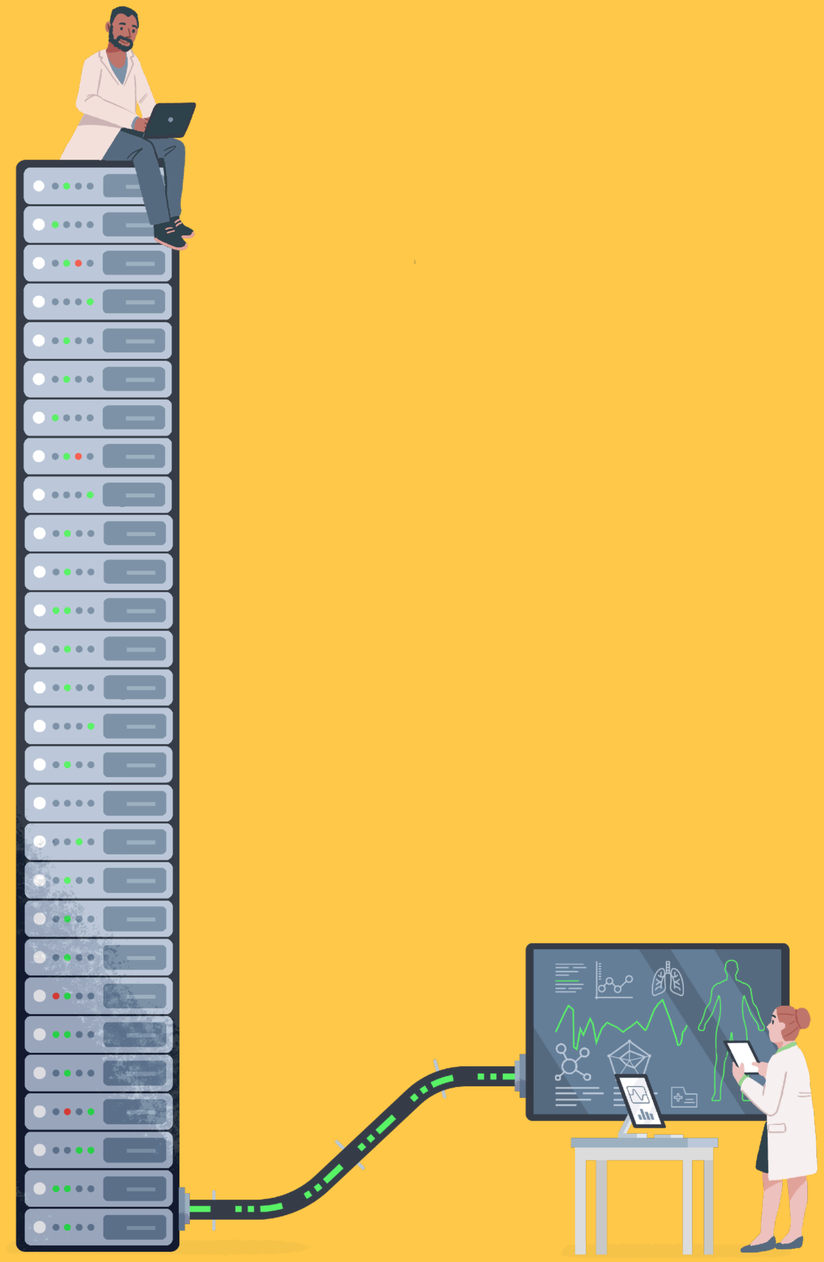
- It is about the '**what?**' of AI assisted decisions
- Informs individuals on what data about them and another sources of data were used in an AI decision
- Also, informs on what data was used to train and test the model
- What **purpose** does this explanation serve?
  - Challenging a decision
  - Providing reassurance



Data explanation

What this type of explanation could include:

- How the data used to train, test, and validate an AI model was managed and utilised from collection through processing and monitoring
- Which data were used in a particular decision and how
- Which training/testing/validating data were collected, sources of that data, and the methods that were used to collect it
- How procured or third-party provided data were vetted
- How data quality, integrity, and security were assessed and the steps that were taken to address any quality, integrity, or security issues discovered, such as completing or removing data



Data explanation

What this type of explanation could include:

- What the training/testing/validating split was and how it was determined
- How data pre-processing, labelling, and augmentation supported the interpretability and explainability of the model
- How potential bias and discrimination in the dataset have been mitigated.

Fairness explanation

- It provides information about the steps taken to ensure AI-supported decisions are **bias-mitigated** and **equitable**
- Gives people understanding of whether they have been treated equitably themselves

- What **purpose** does this explanation serve?
  - Challenging a decision
  - Justified trust

Fairness explanation

This type of explanation includes information about how the following types of fairness have been ensured and corresponding biases mitigated:

- Data fairness
- Application fairness
- Model design and development fairness
- Metric-based fairness
- System implementation fairness
- Ecosystem fairness

Safety and performance explanation

- Provides information about the measures put in place to ensure **accuracy**, **reliability**, **security**, and **robustness** of AI-supported decisions

- What **purpose** does this explanation serve?
  - Challenging a decision
  - Informative purposes
  - Reassurance

Safety and performance explanation

What this type of explanation could include:

For accuracy:

- How accuracy is measured (e.g., maximising precision to reduce the risk of false negatives)
- Why those measures were chosen, and what the assurance process behind it was
- What was done at the data collection stage to ensure that the training data was up-to-date and reflective of the characteristics of the people to whom the results apply
- What kinds of external validation has undertaken to test and confirm your model's 'ground truth'

Safety and performance explanation

What this type of explanation could include:

For accuracy:

- What the overall accuracy rate of the system was at testing stage
- What is done to monitor this (e.g., measuring for concept drift over time)

Safety and performance explanation

What this type of explanation could include:

For reliability:

- How reliability is specified and measured and what the assurance process behind it is.
- Results of the formal verification of the system's programming specifications, i.e., how encoded requirements have been mathematically verified.

For security:

- How it is measured and what the assurance process behind it is, e.g., how limitation have been set on who is able to access the system, when, and how.
- How the security of confidential and private information that is processed in the model has been managed

Safety and performance explanation

What this type of explanation could include:

For security:

- How it is measured and what the assurance process behind it is, e.g., how limitation have been set on who is able to access the system, when, and how

- How the security of confidential and private information that is processed in the model has been managed

Impact explanation

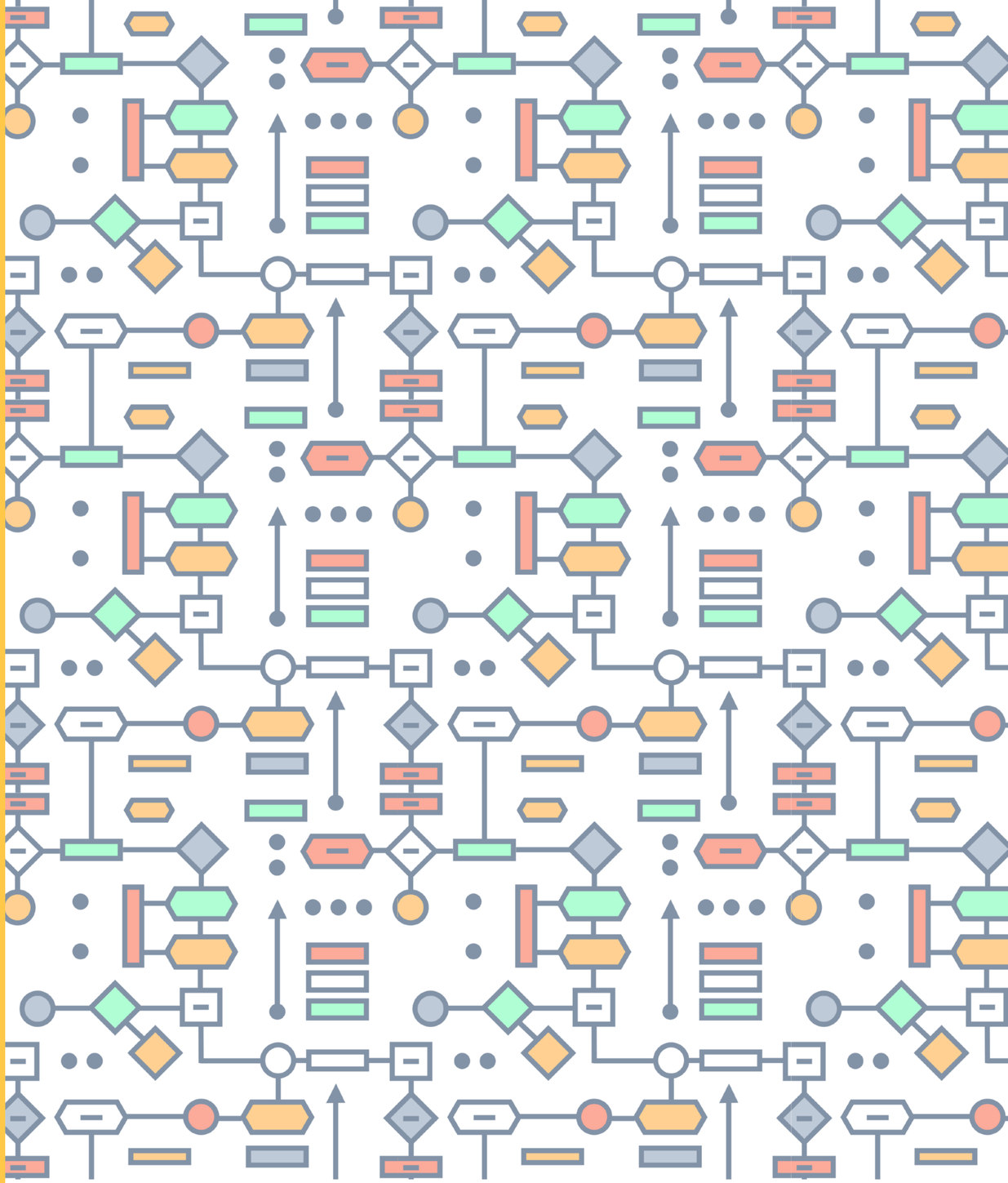
- Provides an explanation on how the **effects** of the AI system on the individual **have been considered**
- It helps people understand the **broader societal effects** that the use of the system may have
- What **purpose** does this explanation serve?
  - Consequences
  - Reassurance

Questions?

Activity 1: Information gathering and evaluating explanations

02

CARE & ACT PRINCIPLES



CARE & Act Principles

CONSIDER CONTEXT

ANTICIPATE IMPACTS

REFLECT ON PURPOSE, POSITIONALITY, AND POWER

ENGAGE INCLUSIVELY

CONSIDER CONTEXT

- No AI system exists in a vacuum.
- Always embedded in a wider socio-technical environment.
- Considering the wider context the system operates in is imperative for responsible research and innovation in AI.

Some questions to bear in mind when considering context:

How are the **norms, values** and **interests of stakeholders influencing** or **steering** the project and its outputs?

How could they influence the users' **meaningful consent** and expectations of **privacy, confidentiality,** and **anonymity**?

How could they shape a **project's reception** across impacted communities?

How are do the **norms and rules** in a domain or jurisdiction shape **expectations** regarding project **goals, practices, and outputs**?

How do the **social, cultural, legal, economic**, and **political environments** influence data generation, the intentions and behaviours of the research subjects, and the space of possible inferences that data analytics, modelling, and simulation can yield?

CONSIDER ...

1/ The contextual determinants of the condition of the production of the project

2/ The context of the users of the system - what are their privacy expectations for example?

3/ The contexts of the social, cultural, legal, economic, and political environments in which projects are embedded as well as the historical, geographic, and other specificities that configure such environments

ANTICIPATE IMPACTS

- Reflect on and assess the **potential short and long-term effects** the system may have

  Impacted **individuals**

  Affected **communities**

- Safeguard the **sustainability** of AI projects **across the entire project lifecycle** instead of dealing with issues as they appear.

- No guarantee that team will anticipate *all* potential impacts

- Dealing with the most relevant ones before they become a problem ensures more sustainable systems overall

ANTICIPATE IMPACTS

- Concerted and purposeful **stakeholder engagement** is essential

- Iterative re-visitation and re-evaluation of s**takeholder impact assessments** secures responsiveness to changing conditions

REFLECT ON PURPOSE, POSITIONALITY, AND POWER

- Scrutiny and reflection on **perspectival limitations** and **power imbalances**
- How do they **influence** the **equity** and **integrity** of the projects and the motivations that steer them?
- **Positionality reflection**
  - Unique experiences and perspectives of all individuals
  - Reflecting on these contextual attributes helps team members understand the gap between their viewpoints and those of stakeholders.

ENGAGE INCLUSIVELY

- Engagement and involve with the community impacted by a project can help **bolster a project's legitimacy** and **democratic governance**
- Ensure that the outputs of process have an appropriate degree of **accountability** and **transparency**
- Always a risk that stakeholder engagement processes are **cosmetic** or **tokenistic**.
- This is why a **proportional** and **responsible approach** to stakeholder engagement is crucial

ACT

ACT transparently and responsibly

Only in engage in AI projects that are both **transparent and accountable by design**.

Ensure that the AI project's governance framework effectively operationalise the values, principles and normative goals that have been prioritised and that safeguard the ethical, equitable and responsible production and use of a system.

Any questions?

Final discussion

Thank you!

For your participation in AI Ethics and Governance