Transparency Regulation Toolkit for Responsible AI

Version 1.1 March 2025

Cite as: Krook J, Winter P, Blockx J, and Downer J, (2024) 'Transparency Regulation Toolkit for Responsible AI,' Responsible AI UK, International Partnerships.

Grant number: EP/Y009800/1

Project

The **RAi UK-funded 'Transparency Regulation Toolkits for Responsible Artificial Intelligence'** project is one of the first empirical studies to explore how Small to Medium-Sized Enterprises (SMEs) are interpreting and implementing AI transparency rules in practice.

The aim of the project was to investigate the practicality of new AI transparency regulations in the UK and EU by hosting stakeholder workshops with industry professionals, policymakers, and academics, fostering dialogue on best practices.

From these workshops, we developed an interactive regulatory toolkit that provides UK and EU SMEs with **practical and straightforward advice** on developing, integrating, and deploying transparent AI systems for the public.

This toolkit is not to be considered legal advice.
For legal advice, please consult a qualified lawyer.

Project Team:

Dr. John Downer
Assoc. Prof. Science & Technology Studies, University of Bristol

Dr. Joshua Krook
Postdoctoral Researcher, University of Antwerp

Dr. Peter Winter
Senior Research Associate, University of Bristol

Dr. Jan Blockx
Asst. Prof. in Law, University of Antwerp

Project website:

https://rai.ac.uk/research/international-partnership-projects/

TABLE OF CONTENTS

01 First Steps

Think about the fundamentals. Who are you trying to be transparent *for*? What are the goals?

02 Techniques

Learn about different transparency techniques including labelling, disclaimers, model cards, explainable AI and more.

03 Culture & Management

Build a culture of transparency in your organization. Make necessary changes to ensure transparency governance.

04 Standards & Certifications

Discover AI transparency standards and certifications, and different organizations working to build Responsible AI.

01

First Steps

First Steps:

This toolkit aims to provide an overview of the many steps you can take to improve AI transparency in your products and services.

To begin, here are a few basic steps you can implement immediately:

1. **Inform your users that your product or service uses AI.**
2. **Clearly communicate any AI-related risks or limitations you know about.**
3. **Add clear and prominent disclaimers specifying how the product should not be used.**

Before going much deeper, however, it is worth reflecting on some fundamentals.

Any transparency efforts will be better if you have a clear sense of their intended purpose and audience.

“What's the purpose of having techniques and tools to be transparent if you don't think about ***why*** you want to be transparent?”

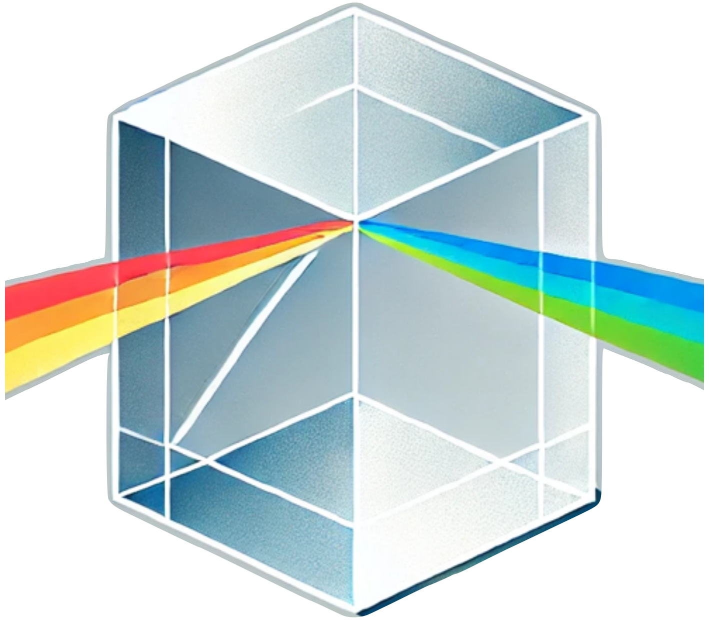
- Bristol workshop participant

Audiences and purposes:

Transparency is inherently **relational**: it's not just about sharing information, but about making it **meaningful** and **accessible** to a specific audience.

Different audiences care about different things and need them to be communicated in different ways.

This means that transparency isn't a one-size-fits-all concept. It's about **understanding who's asking, what they need to know, and how best to communicate it.**



Audiences and purposes:

When addressing AI transparency, it helps to have a sense of your **intended audiences.** These might include:

- **End Users**: Everyday consumers interacting with AI-driven products or services who need to understand its role, limitations, and data practices.
- **Regulators and Policymakers**: Government bodies and agencies framing rules and monitoring compliance.
- **Business Partners and Clients**: Organizations or collaborators relying on your AI systems who expect clear documentation and accountability.
- **Internal Teams**: Decisionmakers within your company needing transparency to ensure responsible design and deployment.
- **Advocacy Groups and NGOs**: Organizations monitoring AI for ethical concerns, bias, etc.

Each group has unique needs and priorities, to which transparency efforts can be tailored.

Audiences and purposes:

Different audiences have unique needs and priorities, to which transparency efforts might respond.

FOR EXAMPLE… If your audience is hoping to use AI transparency to **reduce bias**, then the data you are making transparent must be tailored to the specific bias they're addressing.

- For gender bias in hiring, you might need to reveal performance data across genders to rebalance predictions.
- For racial bias in loans, historical and socioeconomic data could highlight disparities.
- Addressing geographic bias in healthcare might require local data on clinic access and health outcomes.

In these, and a thousand other ways, the **goals of your transparency efforts will shape their implementation**.

Transparency where?

You should also consider what part(s) of the **development cycle** require transparency and **plan accordingly.**

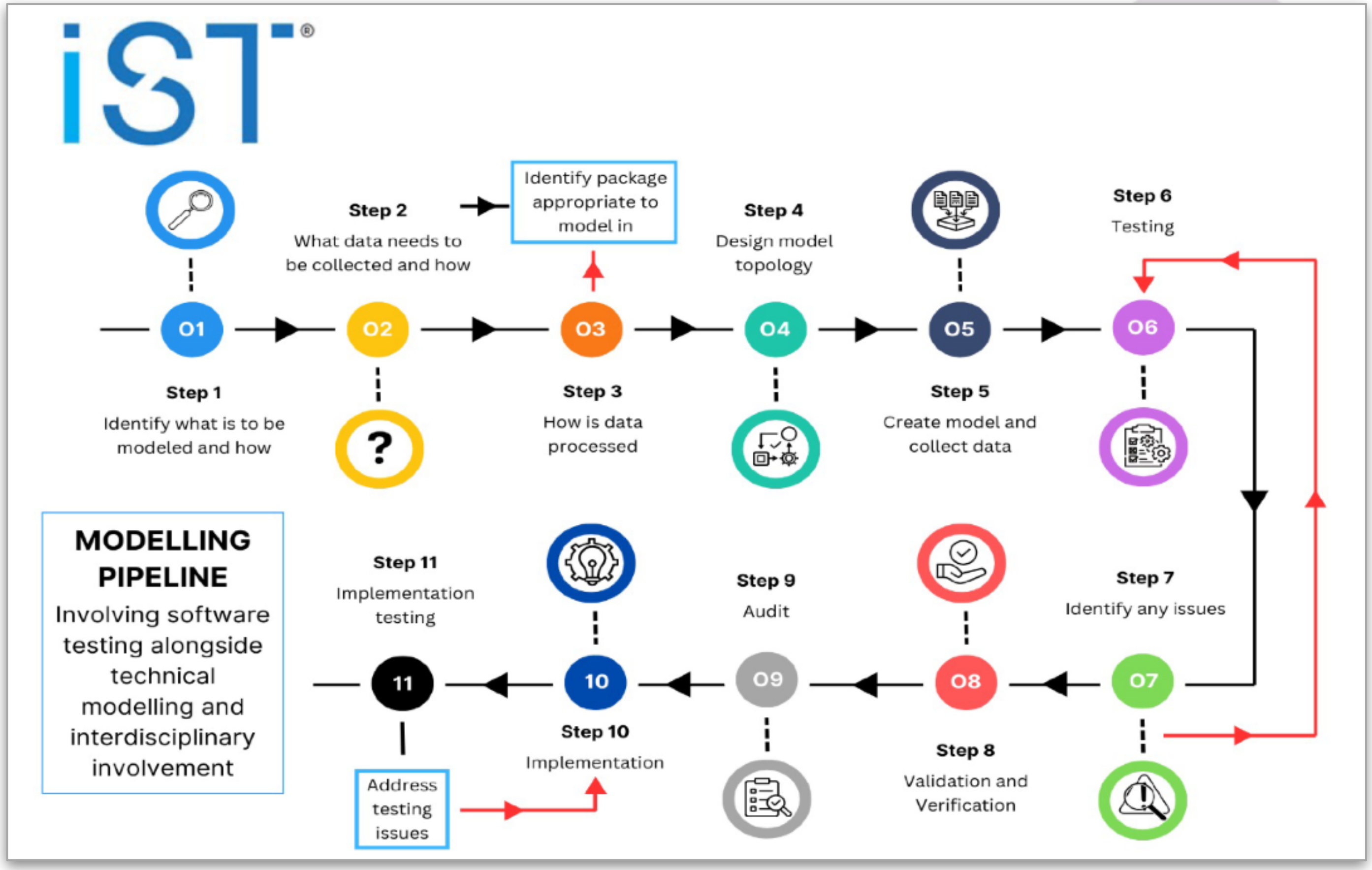

iST®
Step 1
Identify what is to be modeled and how
Step 2
What data needs to be collected and how
Identify package appropriate to model in
Step 3
How is data processed
Step 4
Design model topology
Step 5
Create model and collect data
Step 6
Testing
Step 7
Identify any issues
Step 8
Validation and Verification
Step 9
Audit
Step 10
Implementation
Step 11
Implementation testing
Address testing issues
01
02
03
04
05
06
07
08
09
10
11
?
MODELLING PIPELINE
Involving software testing alongside technical modelling and interdisciplinary involvement

Diagram: The IST Modelling Pipeline shows different stages of the machine learning development lifecycle. At each stage, ask yourself, how can I be more transparent about this process to end-users?

Source: Oldfield, M. (2022) "Towards Pedagogy Supporting Ethics in Modelling," *Journal of Humanistic Mathematics*, 12 (2): 128-159.

02

Transparency Techniques

Transparency Techniques

Labels

Watermarks

Disclaimers

Model Cards

Explainable AI

Data Transparency

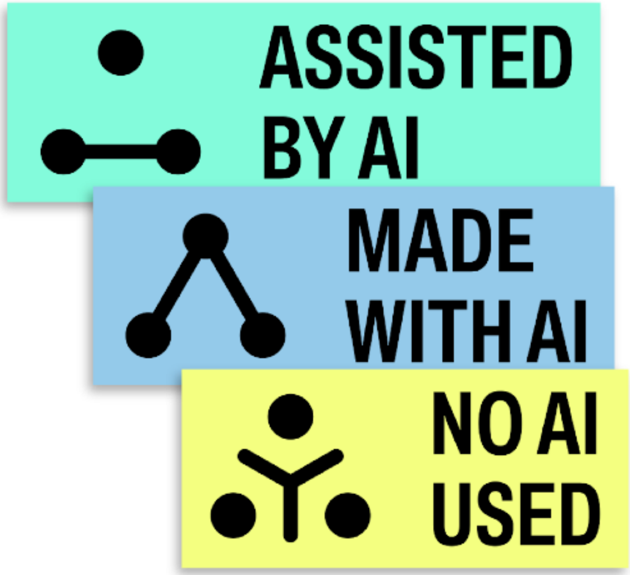
AI Labelling

AI labeling refers to the practice of clearly identifying content or decisions that are generated or influenced by AI systems.

This can include disclaimers, tags, or metadata that inform users about the AI's role, its limitations, and whether the output has been verified or reviewed by humans.

The goal is to increase transparency and help users assess trustworthiness.

Illustrative Implementation:



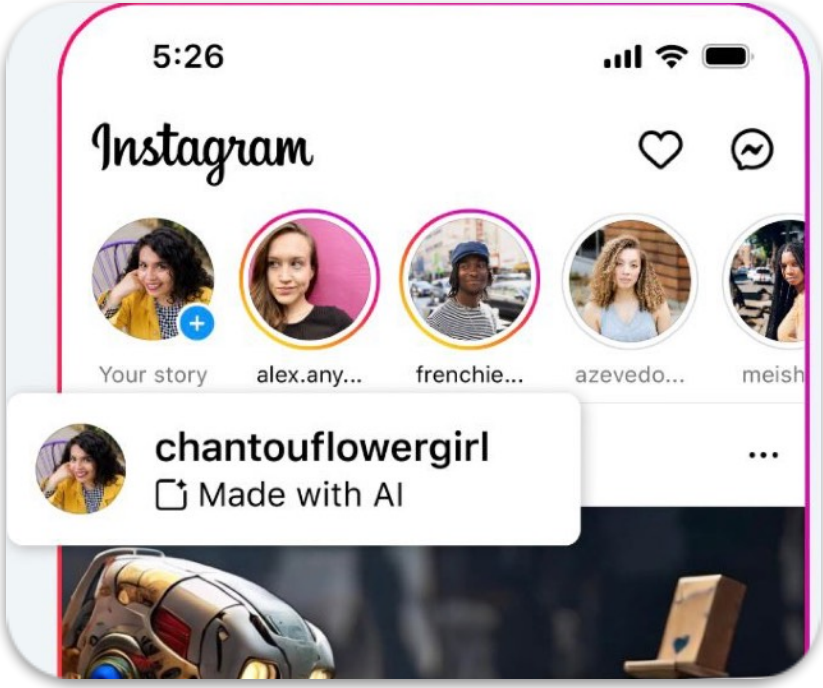
Proposed icons from AI Label

AI Labelling

The Basics:

1. Disclose to users visually when AI is being used in a product or service.
2. Clearly label AI-generated images, audio, video, decision-making or dialog.
3. Consider including further information on data sources and risks.

Illustrative Implementation:



Meta implementing AI labels on Instagram.

Labeling Case Study:
The AI Newsletter

A company creates an AI newsletter, collating summaries of political news for a general readership. However, the AI occasionally generates erroneous and misleading content.

A Simple Solution:

The company implements a new product label, encouraging readers to fact-check their content:

"This newsletter was in part generated and curated by AI agents. They can make mistakes. Please check all necessary information."

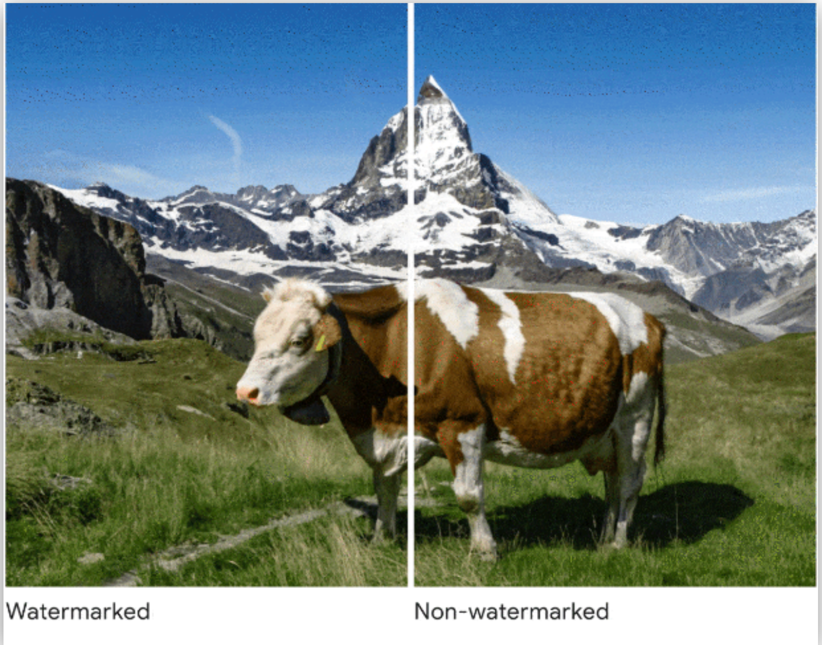
Best Practice:

The company develops a nuanced labeling system that categorizes content based on its source and level of verification. e.g.:

- ***AI-Generated:*** *"This content was generated by AI and may contain inaccuracies. Verified by [Verification Process/Date]."*
- ***Human-Reviewed: "****This content was AI-generated and has been reviewed by a human editor."*
- ***Fully Verified:*** *"This content has been fully verified by our editorial team and aligns with reliable sources."*

AI Watermarking

- AI watermarking embeds subtle, often invisible signals into AI-generated content, like text, images, or videos, to identify it as machine-made.
- These signals don't interfere with how the content looks or feels to humans but can be detected with specialized tools.
- Watermarking enhances accountability by tracking content origins and supports compliance with regulations requiring disclosure of AI-generated material.
- (**Note:** Some watermarks can be removed by third parties or rendered ineffective by advancing technologies.)



Invisible AI watermarking by Google DeepMind.

For further legal information, see: Article 50, AI Act (EU). Also the European Parliament's briefing on "Generative AI and watermarking."

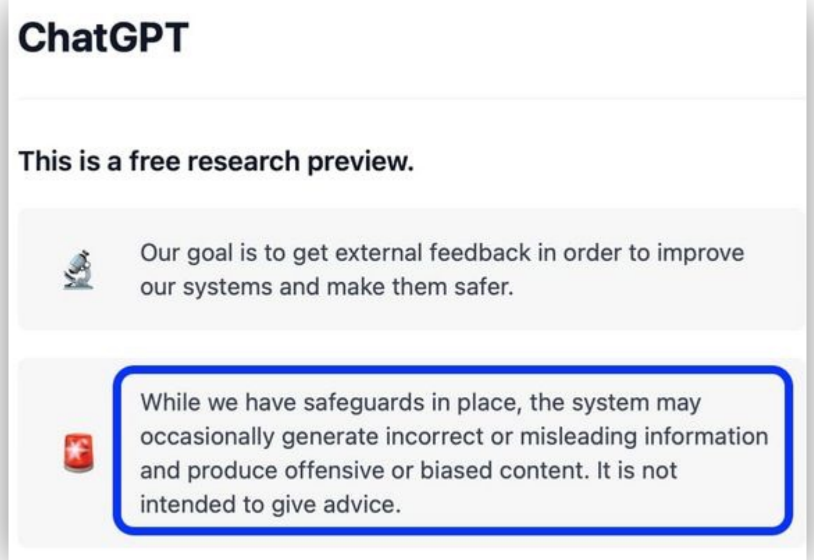
AI Disclaimers

AI disclaimers inform users about the role AI plays in your product or service, its limitations, and how to interpret its outputs.

Here's how to craft effective AI disclaimers:

1. **Clarify AI's Role**: Explain where and how AI is being used. Is it assisting with decisions, automating tasks, or generating content?
2. **Explain Limitations**: Use plain language to describe the risks, potential inaccuracies, and limitations of your AI. Help users understand what it can and cannot do.
3. **Direct to Experts**: Advise users to consult qualified professionals, like doctors or lawyers, for critical or specialized advice that AI isn't designed to provide.
4. **Clarify Intent**: State the AI's intended use clearly, and highlight what it is not meant for.
5. **Establish Accountability**: Provide a risk management plan: identify how users can report errors, or concerns, and specify who will address them.

Illustrative Implementation:



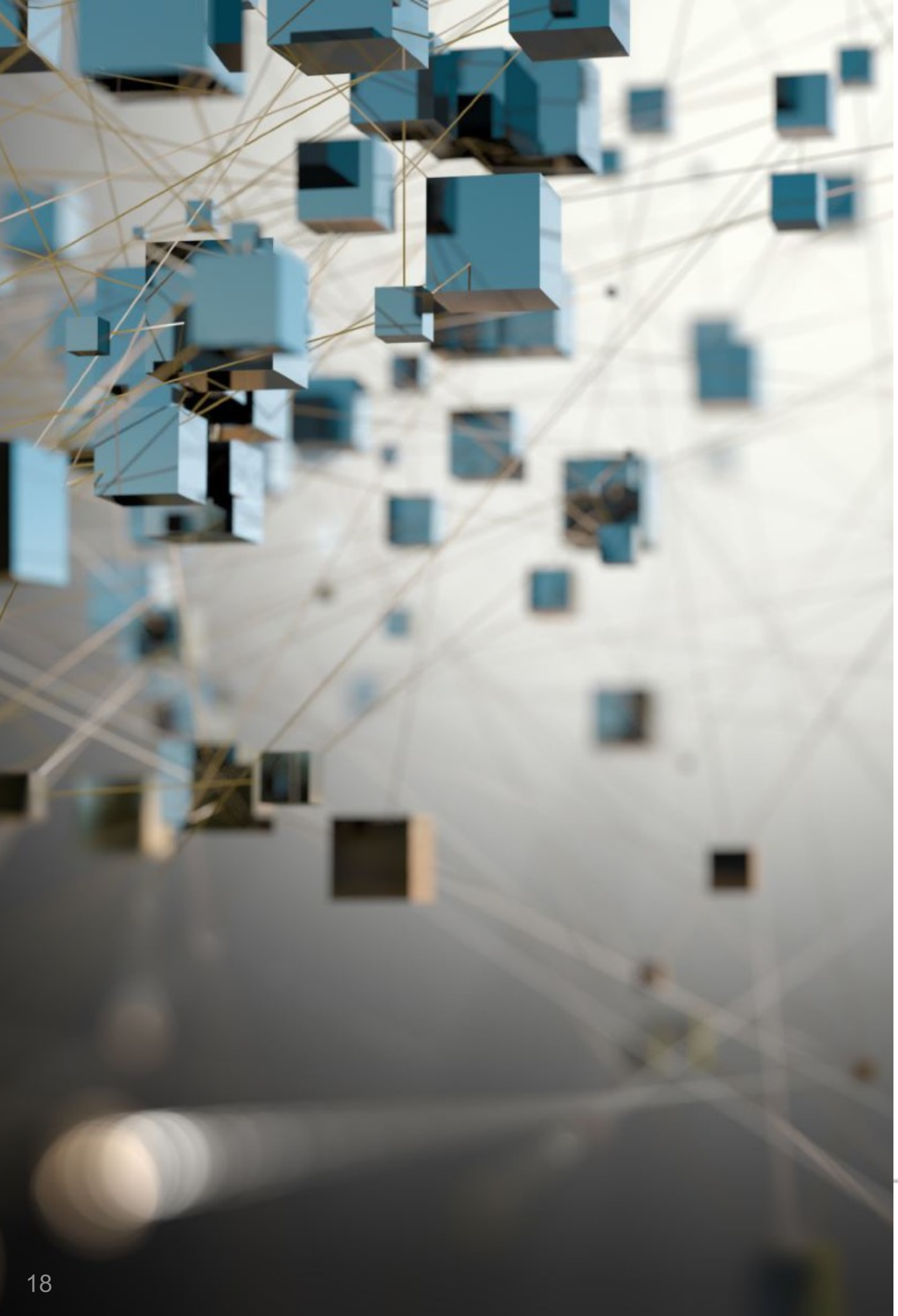
ChatGPT

This is a free research preview.

🔬 Our goal is to get external feedback in order to improve our systems and make them safer.

🚨 While we have safeguards in place, the system may occasionally generate incorrect or misleading information and produce offensive or biased content. It is not intended to give advice.

From: ChatGPT Research.



Data Transparency

“Because we trust it with quite personal information. The amount of people that must have told an AI some of their medical symptoms that they wouldn't want to go and tell their dogs. People are probably trusting these things with a lot, and *what happens to that information*?”

– **Project Manager, SME**

“There's a lot of data everywhere, different sources, fragmented, unstructured, structured. So, one of the problem areas we're looking at is how can we put that all together in a ***structured format,*** … there's the opportunity to tick-off some of the AI safety, AI transparency [rules], where we say ***this is the data we use to train***.’”

– **Data Scientist, SME**

Data Transparency

Data transparency means making your data practices clear, accountable, and accessible to everyone they affect.

Tips for embedding data transparency into your AI use:

1. **Make Data Visible**: Catalog all data your AI uses. Keep records of what you collect, where it comes from, and how it's processed.
2. **Open the Black Box**: Explain, in plain terms, how your AI uses data to make decisions.
3. **Share Sources:** Disclose the origins of your data (whether proprietary, third-party, or public) and any preprocessing steps.
4. **Protect Privacy:** State what personal data you collect, why it's needed, and how it's protected. Always seek informed consent.
5. **Address Biases:** Audit your data regularly for potential biases or blind spots. Explain those steps and their limitations.
6. **Simplify Compliance:** be sure to align with key regulations (like GDPR, CCPA) and ensure users understand their rights.
7. **Invite Scrutiny:** Offer accessible tools or documentation for users to examine your data practices.

Illustrative Implementation:

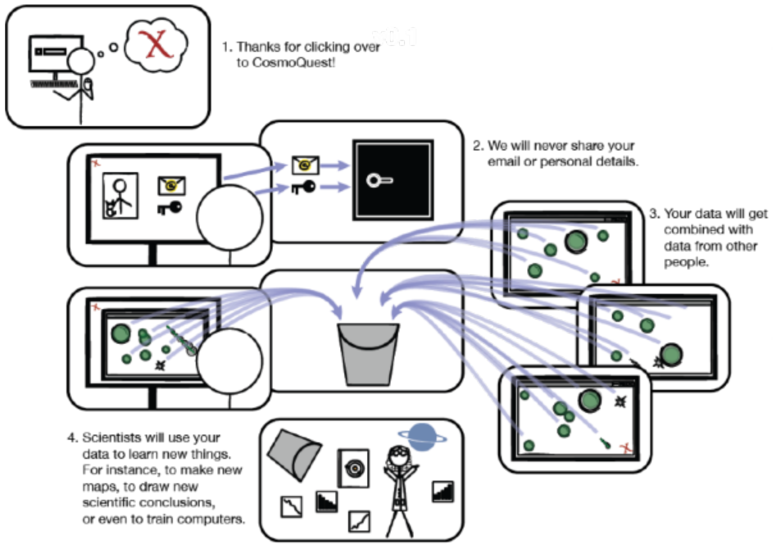

1. Thanks for clicking over to CosmoQuest!
2. We will never share your email or personal details.
3. Your data will get combined with data from other people.
4. Scientists will use your data to learn new things. For instance, to make new maps, to draw new scientific conclusions, or even to train computers.

Diagram: *CosmoQuest* is a citizen science website where users submit recordings, astronomy sightings and other data to advance scientific research.

Model Cards

"So, for example, NVIDIA, they have a lot of models that they provide to people who are using their libraries. They have something called a ***model card***. It's basically some kind of technical limitations that are already known [with] this model. Some extra information about the training data, etc."

- Workshop Participant

Model Cards

AI Model Cards are analogous to food nutrition labels. They provide information about a model's design, purpose and performance in an easy-to-read format.

Key elements typically include:

- **Model details:** name, version, developer, license.
- **Purpose**: What the model is designed to do, and what it isn't.
- **Training Data**: An overview of the data used to train the model.
- **Performance Metrics**: Metrics on the model's performance across different tasks.
- **Limitations**: Details of known weaknesses or risks.
- **Ethical Considerations**: Steps taken to mitigate harm or bias.

Illustrative Implementation:

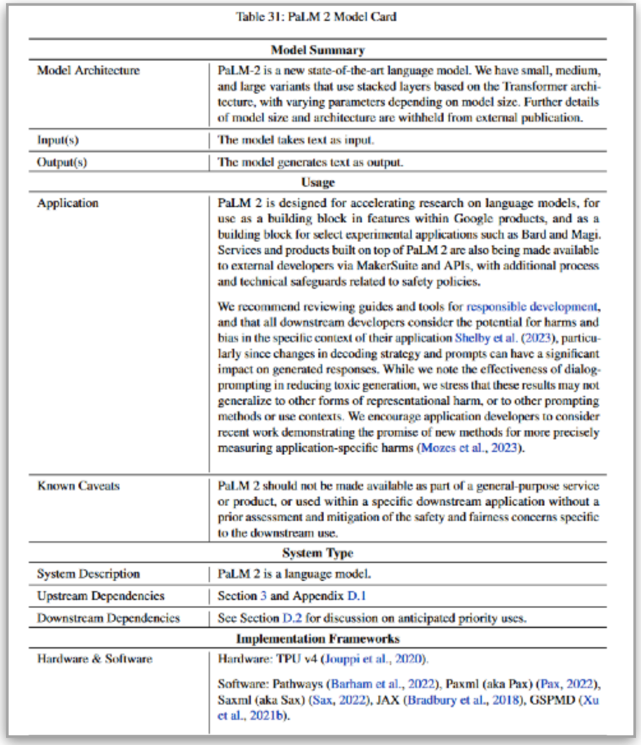


Table 31: PaLM 2 Model Card

Model Summary	
Model Architecture	PaLM-2 is a new state-of-the-art language model. We have small, medium, and large variants that use stacked layers based on the Transformer architecture, with varying parameters depending on model size. Further details of model size and architecture are withheld from external publication.
Input(s)	The model takes text as input.
Output(s)	The model generates text as output.
Usage	
Application	PaLM 2 is designed for accelerating research on language models, for use as a building block in features within Google products, and as a building block for select experimental applications such as Bard and Magi. Services and products built on top of PaLM 2 are also being made available to external developers via MakerSuite and APIs, with additional process and technical safeguards related to safety policies. We recommend reviewing guides and tools for responsible development, and that all downstream developers consider the potential for harms and bias in the specific context of their application Shelby et al. (2023), particularly since changes in decoding strategy and prompts can have a significant impact on generated responses. While we note the effectiveness of dialog-prompting in reducing toxic generation, we stress that these results may not generalize to other forms of representational harm, or to other prompting methods or use contexts. We encourage application developers to consider recent work demonstrating the promise of new methods for more precisely measuring application-specific harms (Mozes et al., 2023).
Known Caveats	PaLM 2 should not be made available as part of a general-purpose service or product, or used within a specific downstream application without a prior assessment and mitigation of the safety and fairness concerns specific to the downstream use.
System Type	
System Description	PaLM 2 is a language model.
Upstream Dependencies	Section 3 and Appendix D.1
Downstream Dependencies	See Section D.2 for discussion on anticipated priority uses.
Implementation Frameworks	
Hardware & Software	Hardware: TPU v4 (Jouppi et al., 2020). Software: Pathways (Barham et al., 2022), Paxml (aka Pax) (Pax, 2022), Saxml (aka Sax) (Sax, 2022), JAX (Bradbury et al., 2018), GSPMD (Xu et al., 2021b).

From: Google (2023) "PaLM 2 Technical Report."

For further information, see Google's model card toolkit.

Explainable AI

“I think it's very important to [consider] the context and a specific use case and a specific user to whom you are being transparent. If you have an AI system where a doctor receives a picture like, okay, your patient probably has a cancer, then “This is assisted by AI” is not enough. Then you will really need to ***explain***, okay, how is it assisted? And why is the system doing it? And it's not an easy quick fix to say, ‘assisted or produced by AI.’ What is AI?”

- **Participant, Antwerp Workshop**

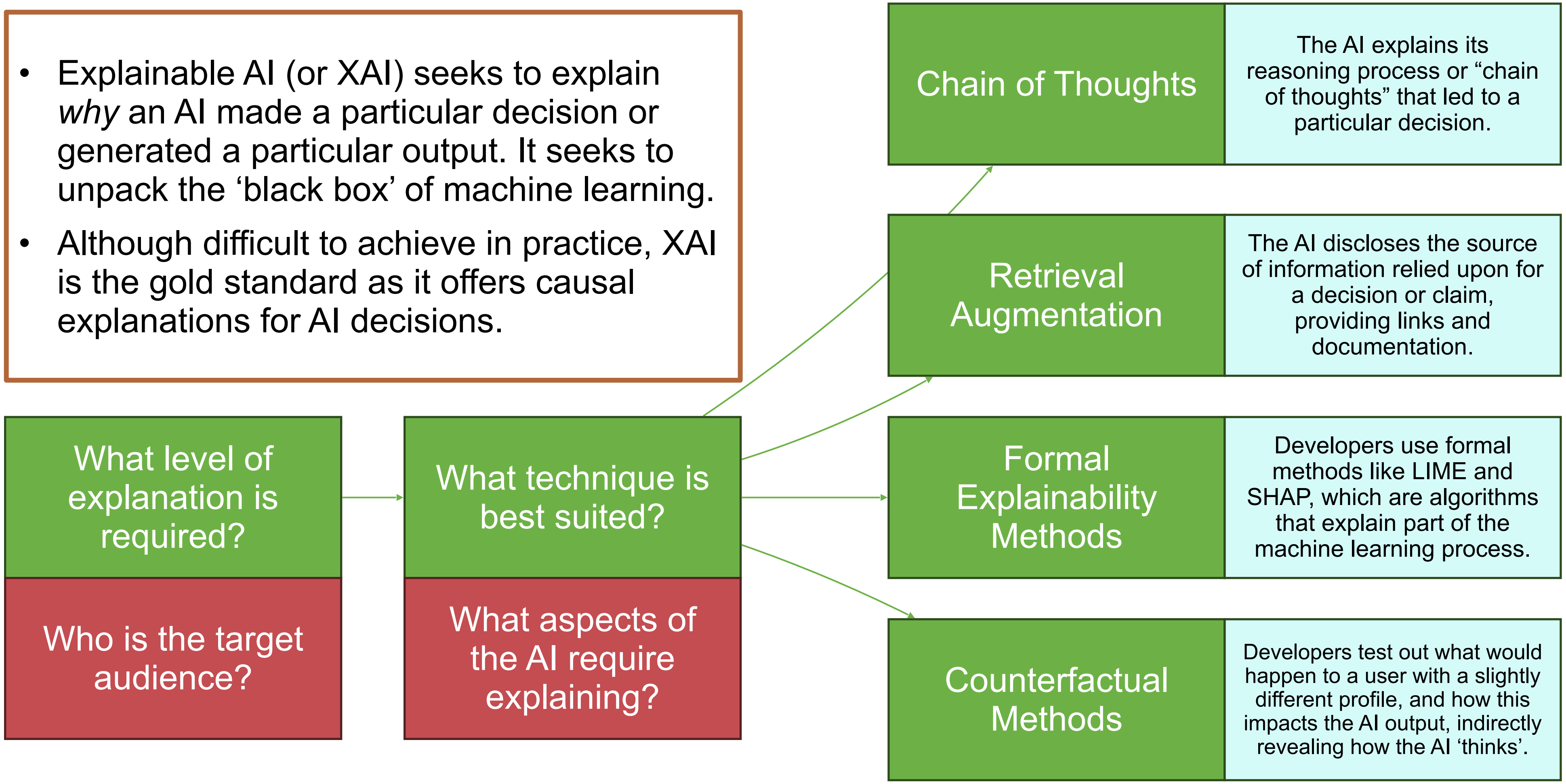
“It could even be counterproductive if you provide too complicated explanations to the end user, then they will not understand it and so they will say, okay, what does this mean? Please provide me with a simple explanation. Why is my loan refused?”

- **Participant, Antwerp Workshop**

Explainable AI

- Explainable AI (or XAI) seeks to explain *why* an AI made a particular decision or generated a particular output. It seeks to unpack the ‘black box’ of machine learning.
- Although difficult to achieve in practice, XAI is the gold standard as it offers causal explanations for AI decisions.

Means of Implementation:



What level of explanation is required?

Who is the target audience?

What technique is best suited?

What aspects of the AI require explaining?

Chain of Thoughts

The AI explains its reasoning process or “chain of thoughts” that led to a particular decision.

Retrieval Augmentation

The AI discloses the source of information relied upon for a decision or claim, providing links and documentation.

Formal Explainability Methods

Developers use formal methods like LIME and SHAP, which are algorithms that explain part of the machine learning process.

Counterfactual Methods

Developers test out what would happen to a user with a slightly different profile, and how this impacts the AI output, indirectly revealing how the AI ‘thinks’.

Case Study:
The AI Diagnostic Tool

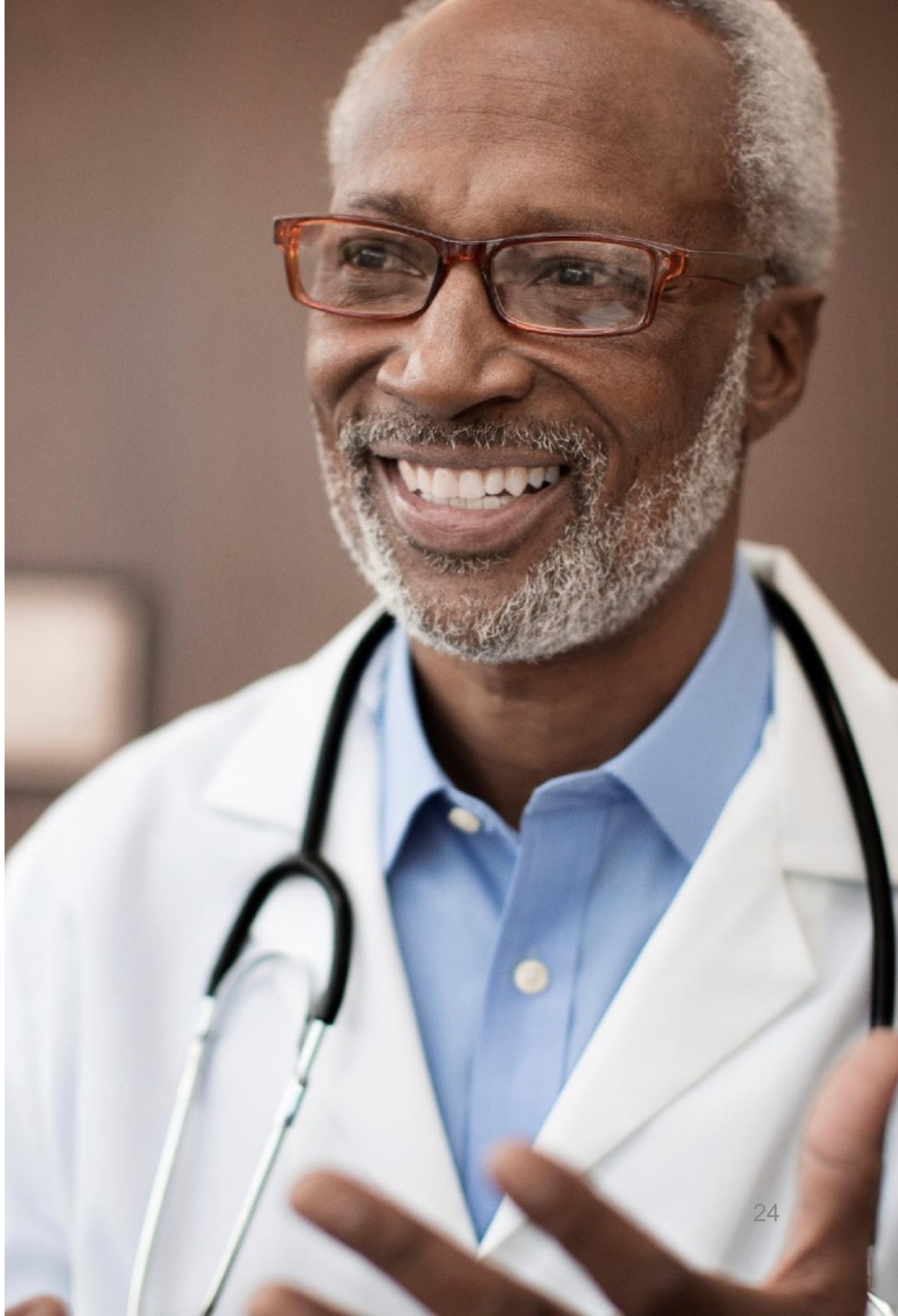
A medical company creates an AI diagnostic tool that gives patients and doctors information regarding test results and a possible diagnosis.

A Simple Solution:

Implement a ‘one-size-fits-all’ AI explainability model that explains the diagnostic decision by revealing what data was relied upon by the AI to reach that decision.

Best Practice:

Implement a more comprehensive AI explainability model that provides different levels of explanation according to the expertise of the end-user, including the potential use of counterfactual methods.



03

Culture & Management

Training and Education

“To come back to the [EU] AI Act, there is an AI literacy obligation as of next year. It's a very short article with a lot of implications for a lot of companies, but there are no requirements on the duration or on the contents.”

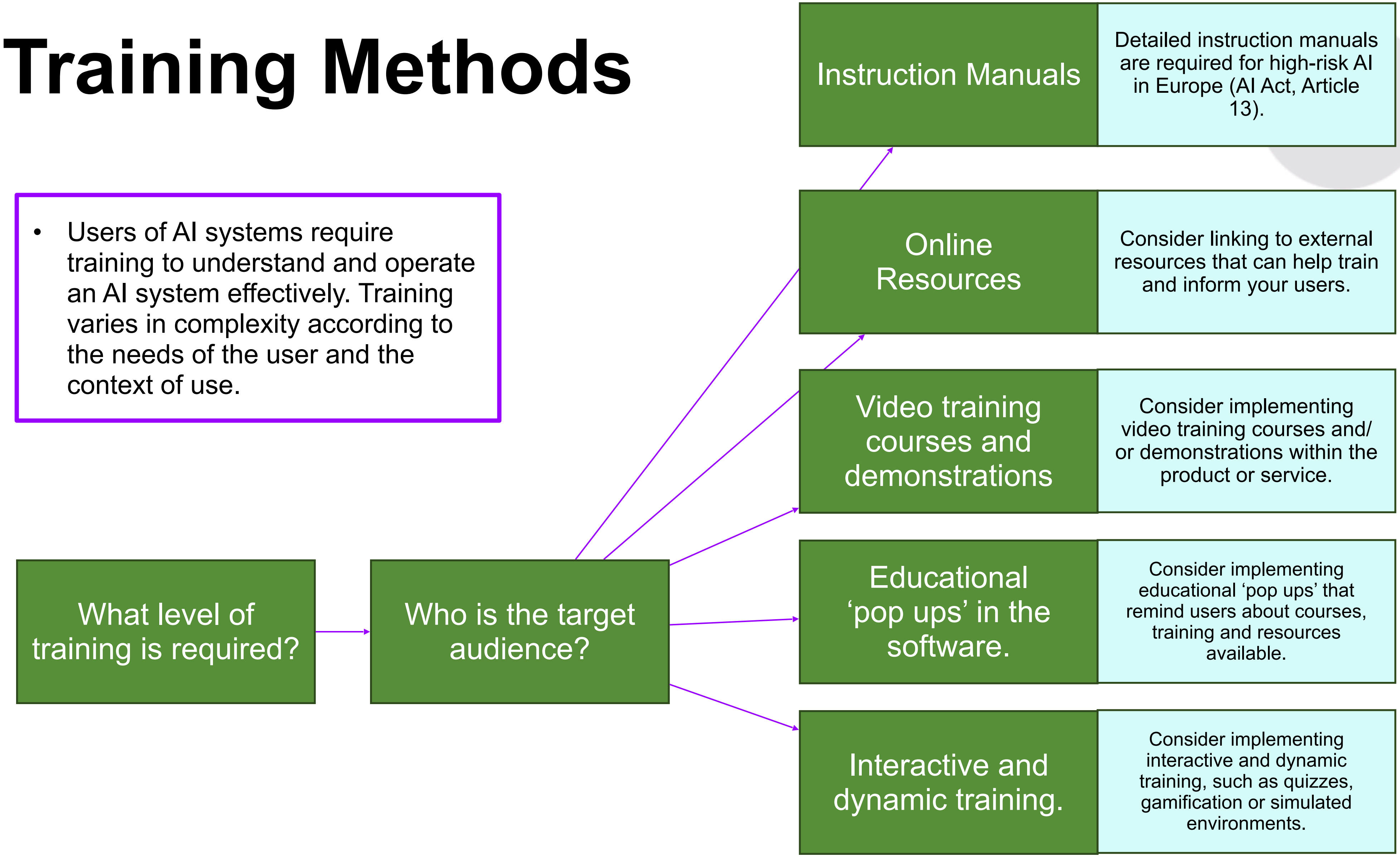
- **Workshop Participant**

“There's a few cases that I know of government agencies forcing their, well, civil servants in this case, to go through training programs to interpret the AI systems that they use, and that was partly driven by legislation.”

- **Workshop Participant**

Training Methods

- Users of AI systems require training to understand and operate an AI system effectively. Training varies in complexity according to the needs of the user and the context of use.



What level of training is required?

Who is the target audience?

Instruction Manuals

Detailed instruction manuals are required for high-risk AI in Europe (AI Act, Article 13).

Online Resources

Consider linking to external resources that can help train and inform your users.

Video training courses and demonstrations

Consider implementing video training courses and/ or demonstrations within the product or service.

Educational 'pop ups' in the software.

Consider implementing educational 'pop ups' that remind users about courses, training and resources available.

Interactive and dynamic training.

Consider implementing interactive and dynamic training, such as quizzes, gamification or simulated environments.

Case Study:
The Maritime Device

A maritime company creates an autonomous shipping device, which presents clear occupational, health and safety risks if used incorrectly, with the potential for harm to the human operator.

A Simple Solution:

Provide a comprehensive instruction manual including all safety features, risks and limitations of the product, including technical specifications, how-to-guides and further resources for users to self-learn at their own pace.

Best Practice:

Provide both an instruction manual and more detailed training materials ***inside*** the application itself. This could include video training, a short course, interactive or synthetic environments or simulations of the maritime environment for users to practice in safety prior to deployment.

(Note: In some cases, this is **the legal standard**.)

Case Study:
The Human Resources (HR) System

A government uses an HR System to assign job seekers to categories which will determine the amount of support they receive. However, this automated system comes with certain risks compared to manual-entry systems.

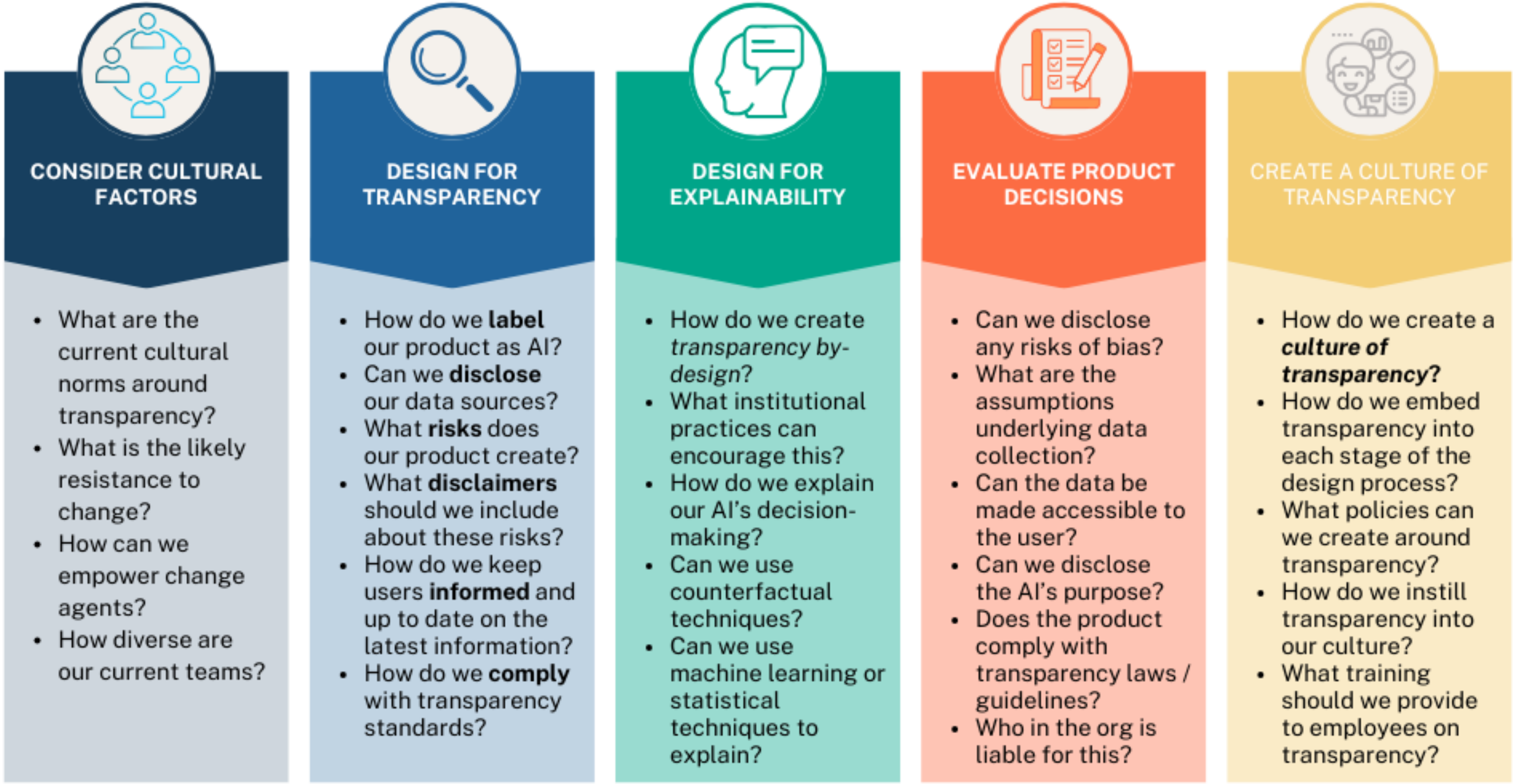
A Simple Solution:

Users receive a disclaimer which tells them to use the system with caution, for it can make inaccurate determinations on their situation and/or result in biased outcomes.

Best Practice:

Users receive training inside of and outside of the product on how best to use the system, the limitations and risks involved in doing so, and how to receive timely support with appropriate grievance and appeal procedures in place.

Governance and Cultural Change



CONSIDER CULTURAL FACTORS

- What are the current cultural norms around transparency?
- What is the likely resistance to change?
- How can we empower change agents?
- How diverse are our current teams?

DESIGN FOR TRANSPARENCY

- How do we **label** our product as AI?
- Can we **disclose** our data sources?
- What **risks** does our product create?
- What **disclaimers** should we include about these risks?
- How do we keep users **informed** and up to date on the latest information?
- How do we **comply** with transparency standards?

DESIGN FOR EXPLAINABILITY

- How do we create *transparency by-design*?
- What institutional practices can encourage this?
- How do we explain our AI's decision-making?
- Can we use counterfactual techniques?
- Can we use machine learning or statistical techniques to explain?

EVALUATE PRODUCT DECISIONS

- Can we disclose any risks of bias?
- What are the assumptions underlying data collection?
- Can the data be made accessible to the user?
- Can we disclose the AI's purpose?
- Does the product comply with transparency laws / guidelines?
- Who in the org is liable for this?

CREATE A CULTURE OF TRANSPARENCY

- How do we create a ***culture of transparency?***
- How do we embed transparency into each stage of the design process?
- What policies can we create around transparency?
- How do we instill transparency into our culture?
- What training should we provide to employees on transparency?

Cultural Factors

An organization’s culture is its personality. This is reflected in its management style, leadership team, workplace policies and employee satisfaction, among other metrics.

Consider:

- What is the current culture regarding transparency?
- How do we empower change agents?
- How do we embed transparency into our practices and policies?
- What training should we provide to employees?

Design for Transparency

Consider:

- How do we **label** our products?
- Can we **disclose** data sources?
- What **risks** does our product create?
- What **disclaimers** should we include?
- How do we create transparency **by-design**?
- How do we **inform** users?
- How do we **comply** with the law?

04

Standards & Certifications

Standards for AI Transparency

AI standards provide technical information on AI Transparency methods and an ethical framework for AI use and deployment.

(1) **International Organisation for Standardisation (ISO)** is developing several AI standards through its subcommittee **ISO/IEC JTC 1/SC 42**. These standards focus on areas such as trustworthiness, transparency, governance, and bias in AI systems.

Relevant work: ISO/IEC TR 24028 (AI Trustworthiness), ISO/IEC TR 24027 (Bias in AI).

(2) **Institute of Electrical and Electronics Engineers (IEEE)** is leading the development of **ethics-focused standards for AI** through the IEEE P7000 series. These standards address transparency, accountability, and fairness in AI and autonomous systems, with specific projects such as IEEE P7001 focused on transparency.

Relevant work: IEEE P7001 (Transparency of Autonomous Systems), IEEE P7003 (Algorithmic Bias Considerations).

(3) **European Committee for Standardisation (CEN) and European Committee for Electrotechnical Standardisation (CENELEC)** are developing **European standards for AI**, particularly in alignment with the EU's AI Act. Their work focuses on ethical guidelines, transparency, and trust in AI systems.

Relevant work: CEN-CENELEC JTC 21 (Artificial Intelligence standardisation efforts).

Standards for AI Transparency (cont.)

AI standards provide technical information on AI Transparency methods and an ethical framework for AI use and deployment.

(4) The **European Office for AI**, through the **European Centre for Algorithmic Transparency (ECAT)**, has been actively working on developing standards for AI transparency. ECAT plays a key role in the implementation of the **Digital Services Act (DSA)**, which aims to ensure that online platforms operate in a transparent, accountable, and predictable manner.

Relevant work: Digital Services Act (DSA) Transparency Requirements, AI Code of Practice, and EU AI Act AI Transparency Standard

(5) **British Standards Institution (BSI)** is involved in **developing ethical AI standards** that cover transparency and trustworthiness. They have published BSI PAS 440 on responsible innovation and are contributing to global AI standards through collaboration with ISO.

Relevant work: BSI PAS 440 (Responsible Innovation), BS 8611 (Ethics in Robotics), and ISO 42001 (AI Management System)

(6) **German Institute for Standardization (DIN)** is actively working on AI transparency standards in collaboration with CEN, CENELEC, and ISO. They focus on ethics and trust in AI systems.

Relevant work: ISO/IEC JTC 1/SC 42 (Global Standards related to AI) and DIN SPEC 92001-1 (Quality Assessment for AI Systems)

Certification

Certification bodies can give your product a ‘tick of approval’ – reassuring customers that you are compliant with relevant rules and procedures.

- **AI Transparency Institute** offers certifications and assessments for AI systems, focusing on transparency, accountability, and fairness. They provide tools and frameworks to help companies make their AI systems more transparent, ensuring the public and regulators can understand how these systems function.
- **European Union** Certification **(Under the EU AI Act)** has provisions for the certification of high-risk AI systems, which include transparency as a core requirement. Systems that meet the EU AI Act's standards for transparency and risk management can be certified for deployment in the EU.
- **IEEE Certified Ethically Aligned Design (CEAD)** has developed standards, including the **P7001 Standard for Transparency of Autonomous Systems**, which can be used as a certification framework. These guidelines help ensure AI systems are designed transparently, and organizations can receive certification based on these principles.
- **BSI (British Standards Institution) Certification** has worked on BSI PAS 440, a guide for responsible innovation, including transparency in AI systems. BSI provides certification services for AI systems that align with ethical and transparency guidelines.
- **AI Ethics Impact Group (Germany)** developed a certification system for ethical AI, focusing on transparency, fairness, and accountability. Their goal is to certify AI systems based on ethical criteria, including transparency in how AI systems operate and make decisions.

Abridged References

- Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). 'On the genealogy of machine learning datasets: A critical history of ImageNet.' *Big Data & Society*, 8(2)
- Department for Science, Innovation and Technology and Office for Artificial Intelligence, (2023) '*AI Regulation: A pro-Innovation Approach*,' Online: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach
- European Parliament and Council (2024) Regulation (EU) 2024/1689 of 13 June 2024 'Artificial Intelligence Act'. Official Journal of the European Union L 327, pp. 1–60.
- Krook, J. Downer, J. Winter, P. and Blockx, J., (2025) "A Systematic Literature Review of Artificial Intelligence (AI) Transparency Laws in the European Union (EU) and United Kingdom (UK): A Socio-Legal Approach to AI Transparency Governance" in *AI & Ethics*
- Oldfield, M. (2022). Towards Pedagogy Supporting Ethics in Modelling. *Journal of Humanistic Mathematics*, 12(2): 128-159.
- von Eschenbach, Warren, J. (2021) "Transparency and the Black Box Problem: Why We Do Not Trust AI." *Philosophy & Technology* 34(4): 1607–22.
- Weston SJ, Ritchie SJ, Rohrer JM, Przybylski AK. (2019) 'Recommendations for Increasing the Transparency of Analysis of Preexisting Data Sets. Advances in Methods and Practices' *Psychological Science*. 2(3): 214-227.

Acknowledgements:

We appreciate the advice and stakeholder outreach provided by:

- The European Digital SME Alliance
- Digital Catapult UK
- The Information Society Law Centre at the University of Milan
- Bristol AI Transparency Workshop participants
- Antwerp AI Transparency Workshop participants