Thinks
Insight & Strategy

Thinks
Insight & Strategy

March 2024

Public Attitudes Towards AI Assurance

Prepared for the Responsible Technology Adoption Unit (RTA) by Thinks Insight and Strategy

Table of Contents

Executive summary

Background

The Responsible Technology Adoption Unit (RTA) within the Department for Science, Innovation and Technology (DSIT) is committed to developing tools that give the public confidence that AI technology works in the way they expect. This is with the aim to build public trust in new technology.

As part of this, the RTA is supporting the development of a globally leading UK AI assurance ecosystem. Assurance is the process of measuring, evaluating and communicating something about a system or process, documentation, a product or an organisation. In the case of AI, assurance measures, evaluates and communicates the trustworthiness of AI systems and their compliance with relevant regulations.

However, the use of AI assurance terminology and concepts is inconsistent within the UK. For example, different organisations can mean different things when using the same term. There are also differences internationally, such as between AI assurance concepts and terminology in the UK and other jurisdictions. Overall, this lack of consistency presents a challenge to communicating and understanding whether AI systems are trustworthy.

Objectives

To support its work on AI assurance, the RTA wants to understand how to communicate about AI assurance concepts. This is with the aim of communicating in a way the public will expect and understand, and through doing so, enable justified public trust. The RTA commissioned Thinks Insight & Strategy (Thinks) to conduct research with the public. The objectives of the research were as follows:

- To understand how well the public understands existing AI assurance terminology.
- To understand how the public thinks actors in the AI assurance ecosystem should talk about AI assurance (at a high level).
- To understand how the public thinks actors in the AI assurance ecosystem should describe specific assurance techniques, in terms of terminology and level of detail.

Methodology

Thinks engaged a total of 35 participants over two phases of research. Each participant took part in two focus groups, delivered a week apart.

- The first focus group explored perceptions of AI technology, trust, and initial views of AI assurance.
- The second focus group explored AI assurance in more detail, including testing certain assurance concepts to understand how they are perceived.

Key findings

The research revealed six key findings:

1. **Participants have a high-level understanding of assurance and associated concepts,** but do not always know how these concepts apply to AI specifically. Understanding of assurance for safety and security is higher than other areas, as participants can draw on a greater number of references from other sectors. On the other hand, fairness is much more complex and less well understood. Participants both question the importance of assuring AI systems for fairness and how this can be done in practice.
2. **The organisation assuring the AI product or service is as important as the process.** Participants want to see assurance delivered by an independent body that has the appropriate level of technical expertise. Participants believe that if assurance were only delivered by an AI developer (that is profit-motivated) then it would not be trustworthy.
3. **Knowledge of how an AI product or service has been assured is not always necessary nor sufficient to build trust.** The context in which participants encounter the application and the perceived risk of an AI product or service are highly influential. Participants do not require knowledge of assurance to trust applications which feel low risk (e.g. facial recognition to unlock a mobile phone). On the other hand, knowledge of assurance is not sufficient to overcome concerns about applications which are perceived as high risk (e.g. self-driving cars). In these instances, many say they would like to see the AI product or service being used by others before they would use them themselves.
4. **Certification as a means of demonstrating AI assurance is well received.** It acts as a shortcut to let participants know a product or service has been reviewed and approved. For most, the knowledge that an AI product or service is certified is sufficient to build trust without the provision of additional information. However as with AI assurance in general, participants want the organisation issuing certification to be independent and expert.
5. **Trust in products and services is reflexive. Participants rely on heuristics such as brand familiarity** rather than a detailed evaluation of testing and governance. These heuristics relieve participants from the additional mental load of evaluating if everyday technology is trustworthy.
6. **Participants easily identify risks of AI technology, focusing on a handful that feel most salient.** Participants are more likely to be concerned by risks created by how AI technology is used, rather than risks inherent to technology itself. For example, participants worry about the risk of widespread job loss, more than the risk of bias.

These findings point to five implications for the AI assurance ecosystem:

1. **It is important to align the communication of AI assurance with other sectors,** given that the public's (limited) understanding of AI assurance is underpinned by references from other sectors.
2. **The public generally do not want – nor need - detailed information to trust an AI system.** Instead, most rely on heuristics, as with other products and services in their lives. This includes certification, highlighting its importance as an efficient way of communicating AI assurance.
3. **Assurance may have greater value for higher risk applications of AI,** where participants want to know more about the detail on how a system has been checked and verified. That said, for some high risk applications there is also a need to consider what other actions (alongside assurance) are required to overcome concerns.
4. **The public want shortcuts, rather than the details. '**Certification' as a means of demonstrating assurance works well to reassure participants that action has been taken and a product or service can be trusted.
5. **Focus on promoting the 'who' rather than the 'what'.** The public feel that the organisation delivering the assurance is as (if not more) important than the specific processes. The organisation should be both competent (i.e. an expert in AI) and have the appropriate motivations (i.e. be independent of profiting from AI technology).

Background and objectives

Assurance is the process of measuring, evaluating and communicating something about a system or process, documentation, a product or an organisation. In the case of AI, assurance measures, evaluates and communicates the trustworthiness of AI systems and their compliance with relevant regulations.

The use of artificial intelligence (AI) assurance terminology and concepts is not consistent within the UK or internationally. This presents a barrier to understanding AI assurance and to consistently and effectively communicating the trustworthiness of AI technology.

To support its work on AI assurance, the Responsible Technology Adoption Unit (RTA) wants to understand how best to communicate to the public about AI assurance. The RTA has commissioned Thinks Insight & Strategy to conduct research with the public on this topic. The research has three aims:

1. To understand how well the public understands existing AI assurance language.

2. To understand how the public thinks actors in the AI assurance ecosystem should talk about AI assurance (at a high level).

3. To understand how the public think Government and industry should describe specific assurance techniques, in terms of terminology and level of detail.

Our approach

What did we do?

We engaged 35 participants over a multi-stage approach:

- **Stage 1:** 6 x 90 minute focus groups.
- **Stage 2:** 6 x 90 minute focus groups.

Both stages were facilitated online, and any associated support participants needed was provided to them.

Stage 1 | Understanding the context and initial views

We covered the following topics as part of stage 1:

- Awareness and perceptions of AI technology, including the risks and benefits.
- Exploring trust and trustworthiness in other sectors, including other technology and financial services.
- Understanding initial views on "AI assurance" and areas of AI assurance (safety, fairness, regulation, accountability).

Stage 2 | Exploring views of AI assurance in more detail

We covered the following topics as part of stage 2:

- Reflecting on the findings from stage 1.
- Exploring understanding and perceptions of AI assurance in relation to four use cases (each with varying levels of risk).
- Exploring views of certification as a means of demonstrating assurance.

Who did we hear from?

We heard from a total of 35 participants across the UK. We recruited participants to achieve:

- **A diverse spread of demographics**, including gender, age, socioeconomic group and ethnicity (detail in the table below).
- **A mix of confidence in and attitudes towards technology** using dimensions included in the RTA's public attitudes segmentation[1]. Dimensions included: digital confidence, awareness of data use, trust in organisations and attitudes towards sharing personal data.

Below is a demographic breakdown of the total sample:

Demographic category	Sub-group	Number of participants

[1] https://www.gov.uk/government/publications/public-attitudes-to-data-and-ai-tracker-survey

	Total	**35**
Gender	Male	16
	Female	19
Ethnicity	Asian or Asian British	6
	Black, Black British, Caribbean, or African	3
	Mixed, or multiple ethnic groups	3
	White British, White European, or White other	23
Socio-economic group	AB	10
	C1	9
	C2	8
	DE	8
Location	Scotland	1
	Midlands	3
	South of England (exc. London)	4
	North of England	14
	East of England	2

Context in which AI assurance lands

Key insights

1. **Most participants are familiar with AI** and see it as part of everyday life.
2. **Participants easily identify risks of AI technology.** However, not all risks identified by participants can be addressed by AI assurance.
3. **Trust in technology and online services is reflexive.** Participants use heuristics (mental shortcuts that help make decisions quickly) to assess if something is trustworthy and therefore safe to interact with.

Familiarity with AI

All participants – even those who are less digitally confident – are familiar with AI. They have typically heard about the technology via word of mouth, the media, and for some, by using it in everyday life (e.g. via chatbots on apps or browsing content suggestions on websites). A handful of more digitally confident participants have knowingly used generative AI tools such as ChatGPT.

"I like the chatbot on the banking apps. I recently used one because I needed a refund on my credit card."

Female, Less Confident Digital User

"I used AI as part of my degree in coding. If there's an error in my code I'll put it in ChatGPT, and it'll tell me what's wrong with it."

Male, More Confident Digital User

Perceptions of AI

Participants see AI as a relatively new technological development. Levels of digital confidence amongst participants impact how they feel about AI. Those who are more digitally confident are typically more excited by the potential of AI, whereas those who are less digitally confident are more nervous about AI technology. These participants often have lower levels of knowledge about both the benefits and risks of AI, and feel more hesitant about trying new technology until it becomes commonplace in society.

Perceived risks of AI

Participants can readily identify a range of risks of using AI technology, but place greater focus on some which feel more salient. The potential psychological or physical harm AI technology may cause is the most salient risk for participants. When asked about their concerns, participants describe that they are very worried by the prospect of 'something going wrong' (e.g. AI technology sharing incorrect information or making the wrong decision) and members of the public

being harmed. This concern is heightened in perceived high-risk settings, such as healthcare and transport.

"The physical and psychological harm. That is worrying…for example if you're using a GP surgery or some sort of medical system where it makes a diagnosis based on the information you're putting in. It's quite alarming, being given the wrong diagnosis for example."

Male, Distrusting Data Sceptic

In addition, the potential negative societal impacts of AI technology are a salient risk for participants. For example, participants are concerned about the increased use of AI leading to job losses in society and a loss of human connection, or about the use of AI technology by bad actors to spread mass misinformation. These broader societal impacts sit outside of the scope of AI assurance.

"There are enough unemployed people and if this increases because of AI then we need to start rethinking things."

Female, Less Confident Digital User

Other risks inherent to the design and safety of the technology, such as bias – which is within the scope of AI assurance – are less salient. Apart from some participants from ethnic minority backgrounds and those who feel more confident in how AI works, participants are mostly unaware of the risk of bias. Most participants struggled to understand how technology could generate unfair outcomes in e.g. job application processes.

"AI is a programming language. So the inherent biases of people programming are sort of manifesting itself there."

Male, More Confident Digital User

Perceived benefits of AI

All participants identify speed and efficiency as the key benefits of AI. Participants feel these benefits apply in their day-to-day life and specialist settings. Participants feel the speed and efficiency of AI will have the biggest benefit in science and medicine, where it can deliver important breakthroughs at a greater speed than humans. However, they also see the benefits of AI in optimising daily life, for example as a communication or home management tool.

"I can use my Amazon device to switch things on and off vs. actually having to physically do it myself so that's a real benefit, and that I can control heating in my house when I am out and about. It really benefits my physical disability."

Male, Distrusting Data Sceptic

A small number of participants see AI as a tool to help to improve public safety (e.g. by analysing CCTV footage) and drive economic growth.

Why do participants perceive other products and services as 'trusted'?

In order to draw lessons for AI assurance, we asked participants to think about what made them trust other products and services. This included the criteria they look out for (e.g. that it works, that it is safe) and how they know products and services deliver on those criteria.

The discussion demonstrated that participants' trust in products and services is reflexive. For the most part, they do not have the time or motivation to engage in the mental load of evaluating if a product or service is trustworthy before they use it. To this point, they do not ask questions such as 'will it work?' or 'is it safe?' before deciding to use most products and services.

Instead, participants rely on a series of heuristics to determine the trustworthiness of products and services. These include:

- **Brand name.** If a product or service is made by a large or well-known brand, participants see this as an indication that the service or product offered can be trusted.
- **Trusted word of mouth.** Participants look to see if their friends, family and other people they know are using similar products and services. If a product or service is normalised, they see it as indicating it is trustworthy. A recommendation or seeing someone else have a positive experience can also enhance trust.
- **The look and feel** of a product or service. If a product or website looks well made (e.g. high resolution, clear language) then that indicates they can trust it.
- **Certification.** Symbols such as the security padlock on websites and the FCA watermark serve as shortcuts to participants to show that products and services have been checked and approved and are therefore safe to be used.
- **Location of a company.** Participants feel more likely to trust companies with a strong commercial presence in the UK or European Union as they know there are laws and standards that products or services must meet.

"If I'm buying a new phone, if I bought something from Apple I'd expect it to be good. If it's a brand I trust I wouldn't even have it in my mind that I wouldn't be able to trust it."

Male, More Confident Digital User

"If a website doesn't have a padlock on it, it's not secure so I wouldn't go on it."

Female, Distrusting Data Sceptic

Whilst participants assume that checks or tests take place (e.g. performance testing, audits, regulation that needs to be followed), they have little appetite for knowing the detail of what measures, checks or rules are being followed.

AI assurance

Key insights

1. **Participants understand AI assurance concepts and terminology at a surface level.** Understanding is based on participants' somewhat limited experiences with assurance in other sectors, such as financial regulation and product testing. This allows them to imagine the kind of testing and inspection AI might be subject to.
2. **However, participants do not currently look for assurance measures to evaluate if an AI product or service is trustworthy.** Instead, they rely on heuristics (as discussed in the previous section).
3. **There are several factors which influence how reassuring (or otherwise) AI assurance is.** These include existing perceptions of the risk of AI; perceptions of the risk of the given use case; the perceived role of AI in society; and *who* is assuring the AI product or service.
4. **In this context, assurance for safety, security and robustness is well understood and feels reassuring.** It ties into concepts participants are already familiar with (e.g. product testing) and directly addresses a key risk of AI, namely the risk of physical or psychological harm.
5. **On the other hand, assurance for fairness is less well understood and does not feel reassuring.** Participants struggle to draw examples of testing for fairness in other sectors. Plus, most do not see bias as a risk of AI products and services.

Understanding of AI assurance

Participants have not heard of the term "AI assurance" before. However, familiarity with other examples of assurance from other sectors means they can understand the basic principles quickly. For example, participants draw on:

- **Product testing,** particularly car safety testing. Participants understand that AI technology will be tested to make sure it works and is safe in a controlled environment before being released for public use.
- **Regulatory compliance.** Participants draw on examples from other regulatory bodies, such as Ofsted or the FCA, and assume that AI products will also be 'inspected' to ensure they are safe. However, this assumption is vague: participants do not specify what they want inspections to focus on nor specifically who they want to carry them out.

"When you make the [loan] application there is typically information as to how they act in accordance to the FCA and they have to comply with that, so that provides you with trust."

Male, More Confident Digital User

However, beyond this, participants lack a more detailed understanding of the intricacies of AI assurance and associated terminology.

Perceptions of AI assurance

There are several factors which influence participants' perceptions of AI assurance concepts and terminology.

- **Participants' perceptions of the risks of AI technology in general.** The risks of physical and psychological harm are more salient than risks to security, privacy and bias. This may be why participants are more drawn to – and reassured by – assurance to manage the safety and robustness of AI products. On the other hand, fairness, regulatory compliance and accountability are less resonant.

- **The perceived level of risk of the AI use case.** The context in which participants will encounter AI assurance is important. For use cases which are perceived as low risk (e.g. facial recognition software on mobile phones), knowledge and understanding of how it's been assured is not necessary for it to be trusted. On the other hand, knowledge and understanding of how riskier use cases (e.g. self-driving vehicles) have been assured is not always sufficient to build trust. In the case of the latter, *knowing* that cars have been safety tested does not mean participants believe they're safe.

> *"It's all the stuff you want to hear but at the end of the day it's tech...you could be driving along the motorway and you get a glitch, you'd be stuck."*
>
> *Male, More Confident Digital User*

The perceived role of AI in society. Tied to both factors above, in some cases participants do not agree with certain uses of AI technology. This means they automatically view these applications with mistrust and therefore find assurance less reassuring. This is often due to the risk they pose to job loss or human connection, for example, AI healthcare assistants or the use of AI in education.

> *"It would be a shame if AI replaced face to face appointments and being able to speak to a consultant, I think it just adds to the loneliness of interacting with technology."*
>
> *Female, Less Confident Digital User*

- ***Who* is assuring the AI product or service.** Participants believe that AI must be assured by a credible organisation otherwise it will be ineffective. In particular, participants do not believe the developers of AI technology are in a position to evaluate their own products. They feel that these companies are profit motivated and therefore have a vested interest to

clear applications for public use. Instead, there is support for an independent body that has the required expertise and no ulterior motive.

"I'm troubled by the lack of transparency in the whole framework. Who is accountable when it goes wrong? We need a body we can trust, not the makers of this stuff."

Female, Less Confident Digital User

- **A perception that development is outpacing governance.** Participants feel the AI industry is a 'Wild West' due to the current lack of governance and rapid pace of continued development. This means some participants questioned how effective AI assurance will be as it will be playing 'catch up'.

"There's no route to question if you need to challenge something. At the moment it seems like the Wild West, there are no controls."

Male, More Confident Digital User

Understanding and perceptions of specific AI assurance concepts and terminology

Over the two rounds of focus groups, we tested terminology relating to AI assurance. These were introduced in the first focus group, divided into the following four areas:

- Safety, security and robustness.
- Accountability.
- Fairness.
- Regulatory compliance.

In the second focus groups, these terms were tested within the context of specific AI use cases. For example, exploring perceptions of 'performance testing' in regards to self-driving vehicles.

The AI use cases we tested are:

- Personalised healthcare.
- Designing lesson plans in education.
- Facial recognition for unlocking mobile phone.
- Driverless cars.

The reactions to the use cases are described in the blue boxes. For each term we explored:

- Level of understanding and level of assurance.
- Perceived strengths and weaknesses.

- Key takeaways.

Safety, security and robustness

All participants find assurance for safety easy to understand and think it should be mandatory: the idea that AI will undergo safety and security checks before being used in society is important to participants. However, specific concepts are less well understood. In this context, participants feel that terminology that is easy to understand and talks directly to the public's concerns will be the most effective.

It should be noted, however, that requests for more detail on assurance terms came after detailed discussion, rather than as an unprompted reaction. While participants say they want more information on many of the areas of AI assurance, they already use some products or services like Alexa or facial recognition without needing it.

Term tested: Performance Testing	
Level of assurance	**Medium**
Understanding	**The term is clear:** participants know it from other sectors or products and can easily apply it to an AI context.
Perceived Strengths	• **Testing is a popular concept:** participants like to think the AI they will use has undergone several rounds of testing to ensure it is safe and suitable to use.
Perceived Weaknesses	• **AI creators benefit from performance testing:** participants feel that performance testing is something AI companies would do to improve their product. Their interests do not necessarily align with the public's. • **Performance testing does not mean something has been verified as safe:** improving performance may make AI more efficient, or more powerful, but it doesn't necessarily protect the public from its risks.

Key Takeaway	Participants like the idea of testing but it needs to be explicit what is being tested and why. They believe that 'performance' denotes improving AI, most probably for its creator's sake, rather than the public's safety.
Evidence	*"Performance testing sounds more like things running smoothly as opposed to data security."* *Female, Distrusting Data Sceptic*

Term tested: Formal verification	
Level of assurance	**Low**
Understanding	**The term feels like jargon**: Many find this term hard to understand and find the explanation insufficient.
Perceived Strengths	• **'Formal' denotes a proper process:** participants like to know that an official process is in place, reducing the risk of unreliable AI making it through to public consumption. • **It suggests a rigorous check has been carried out**: verification suggests a check has been made thoroughly.
Perceived Weaknesses	• **Participants aren't sure how safety would be 'verified':** this term feels vague to many, they feel that verifying safety doesn't provide enough information about the check, in contrast to a term like 'testing'. • **It feels pointless in the context of cars (the application tested):** some point out that all cars need to formally verified. They question what is different about it in the context of AI, implying that there should be a

	more thorough check given the lack of human driver.
Key Takeaway	Participants need more explanation attached to this term, including more detail about the process of formal verification. With this included, however, it can provide reassurance as participants like the official process and thorough checks implied.
Evidence	*"They're saying something but it means nothing…What form does formal verification take? I need more detail."* *Male, More Confident Digital User*
Term tested: Validation	
Level of assurance	**High**
Understanding	**Language is official but clear:** in contrast to the terms described as jargon, participants feel 'validation' has a clear meaning.
Perceived Strengths	• **Participants feel this term denotes independence:** this term is popular because its explanation includes reference to the check being independent, something participants think is important to properly assess AI. • **It denotes a formal process:** participants like that the process of validation feels official, implying that checks are carried out to a high standard and would catch serious risks to the public.
Perceived Weaknesses	• **More detail requested:** some feel that they would need more information about the process of validation in order to be fully reassured that this process is a thorough check.

Key Takeaway	Validation can be an effective way of letting people know that AI is being checked as part of a formal process of review. Including more detail on the validation process can make it more reassuring for those who want more information.
Evidence	*"What type of validation? Who are the relevant authorities?"* *Female, Less Confident Digital User*

Term tested: Safety testing	
Level of assurance	**High**
Understanding	**It is clear and easy to understand:** the concept of safety testing is easy for all to understand.
Perceived Strengths	• **This term speaks to key risk:** physical safety, especially for a system like driverless cars, is easy to understand and visualise, compared to bias audit in education. • **It is tangible:** participants liken it to practical safety, such as seat belts and airbags. This makes it easier to visualise and think about in practise.
Perceived Weaknesses	• **Some question what the threshold for safe is:** this criticism is more of a reflection of doubts about driverless cars than the term itself.
Key Takeaway	This term's effectiveness comes from its clarity and easiness to understand, while speaking directly to a clear risk of an AI system. It could be improved by including safety, a suggestion several participants make directly. For some though, they would need to see more than testing to get into a driverless car, highlighting the importance of context.

Evidence	*"Safety testing means seat belts, air bags…those sorts of things."* *Male, More Confident Digital User*

Use Case: Self-driving cars

Perceived risk: High

How this use case affects views towards assurance terms: the high perceived risk of this use case dominates discussions among participants. Those who are confident in technology are doubtful about its safety even if it has been through thorough testing and evaluation. However, as this testing is more intuitive to understand, participants are able to understand assurance terminology clearly in this context. Many refer to safety testing in cars now, giving them a useful reference point. Ultimately, while assurance can help provide confidence for some, others feel that regardless of the terms provided, they would only consider using a driverless car once they were in wide circulation in society and proved to be safe by extensive public use.

Fairness

The area of fairness is much more complex for participants, in contrast to safety. Its importance is questioned by some, while others struggle to see how AI can be assured for fairness.

Even among those who understand the risk of bias, assurance for fairness provides little reassurance. Many don't understand how prejudices can be prevented in computer programming when they are so prevalent in humans. They question how an imperfect human can be trusted to assess the fairness of AI, especially when one person's idea of fairness differs from another. To fully understand its importance, the public requires more detailed explanation and clarification.

As mentioned for safety, security and robustness, requests for more detail and information should be viewed within the context of the research and the use of stimulus.

Term tested: Bias

Level of assurance	**Low**
Understanding	**Participants are familiar with the term**: they understand what bias is and how it would affect people, especially in the context of education. However, there is less understanding about bias in AI technology.
Perceived Strengths	• **Bias is a key issue for some**: a group of participants feel that giving everyone fair treatment is an important and relevant issue for today, and appreciate that this is being considered among other risks.
Perceived Weaknesses	• **Checking bias in AI raises questions:** even among those who think it is important, participants are struggling to understand how AI can be assessed for bias, especially as the humans who program it will also be subject to bias. • **Others do not prioritise it as a risk**: some feel that bias is not one of the key issues or risks related to AI, especially in the context of the use cases shown. They are preoccupied with other doubts they have about the use of AI in such circumstances.
Key Takeaway	The issue of bias in AI will need to include a clear explanation about how evaluation can measure this, while also showing why it is important for those who deprioritise it.
Evidence	*"With bias, it comes down to the morals that someone has. How do you put that in AI and how do you measure that?"* *Male, Distrusting Data Sceptic*

Term tested: Bias Audit	
Level of assurance	**Low**
Understanding	**The term feels like jargon**: many are unclear what an audit will entail when it pertains to bias.
Perceived Strengths	• **Audit means thorough, regular checks:** some appreciate the use of the word audit because it implies a vigorous review of an AI system, one that is repeated regularly.

	• **Bias is a key issue for some:** as mentioned, some are pleased to see evaluation addressing an important issue.
Perceived Weaknesses	• **The concept is difficult for many to understand:** Many are disengaged by the use of what feels like a technical term.
Key Takeaway	When discussing fairness, bias or discrimination, participants need clear language to engage. They feel that bias audit is too unclear in meaning, with the terms 'bias' and 'audit' an unusual combination, making it sound jargonistic. They therefore do not believe it is an effective way of communicating this concept.
Evidence	*"The bias audit sounds like a checklist, is it done pre or post? It needs to be a lot clearer."* *Male, Distrusting Data Sceptic*
Term tested: Discrimination	
Level of assurance	**Low**
Understanding	**It is understood by all**: participants know what discrimination means and why it is something that needs to be addressed. However, participants are less sure what it means in the case of AI.
Perceived Strengths	• **Discrimination is a key issue for some:** as with bias, some are pleased to see evaluation addressing an important issue.
Perceived Weaknesses	• **Raises questions about how this can be done:** again, as with bias, most question how effectively AI systems can be checked against discrimination, when this issue hasn't been addressed properly in society. • **It feels vague in the context of AI:** those who are more engaged in fairness as an issue want more details on exactly how discrimination will be prevented in AI.
Key Takeaway	This term, as with bias, is harder for participants to understand, especially with some placing less importance on it and others feeling it is subjective. A clear explanation, possibly using a more detailed example, would help provide more clarity to the public.

Evidence	*"How much data are you having to give them so they're not discriminating against you? A child's race and gender doesn't have relevance to their learning."* *Female, Distrusting Data Sceptic*

Use Case: Designing lesson plans in education

Perceived risk: Medium

How this use case affects views towards assurance terms: similar to the use of AI in healthcare, this AI use case feels more controversial and participants get naturally hung up on the debate about whether it should be introduced at all. They see it as an example of the replacement or loss of human function, which most view through a negative lens. Their focus on assurance terminology therefore centres on whether the technology works in improving education at all, rather than fairness which appears to be less of an immediate priority, even for those who see it as an important issue.

Once the conversation is on fairness in education, the debate on this subject and variance on views mean that explicit explanation is needed to focus participants on what these terms mean in practise.

Accountability

Participants react positively to the terms tested within this area and feel they are an important aspect of AI assurance. However, in some cases participants did not understand the terms correctly. This was especially true for the term 'governance' which participants often interpreted as external governance (e.g. legislation) rather than internal corporate governance.

In the abstract, the terms provide a good level of reassurance. Participants are keen to see a robust system of checks and balances. However, they are less reassuring if they are solely internal corporate assessments. This is because participants feel AI developers should not be the only organisations reviewing their products as they are profit motivated.

Term tested: Risk assessment	
Level of assurance	**High**
Understanding	**It is clear and easy to understand**: participants intuitively understand the role of risk assessments in

	assessing AI. Many have come across it in a different context in their professional life.
Perceived Strengths	• **It speaks directly to participant concerns:** participants see why a risk assessment is necessary for new AI products and services. They feel that it could prevent AI that poses a physical or online threat from being released.
Perceived Weaknesses	• **The term doesn't specify the risk it is assessing**: participants are unsure which risk it will be used to assess, especially given the education context, i.e. participants question whether it will be assessing the consequences of replacing a human function with a computer. • **Participants question who will carry this out:** participants question who will be carrying out the risk assessment.
Key Takeaway	'Risk assessment' works well in providing reassurance because it's well understood and directly addresses risk. Adding specific details, such as who is checking and what they are checking for, can help make it even stronger.
Evidence	*"I feel like risk assessment is a good term, because every risk will be considered and then they will put steps in place."* *Female, More Confident Digital User*

Term tested: Impact Assessment	
Level of assurance	**Medium**
Understanding	**It is clear and easy to understand**: participants understand the role of an impact assessment in assessing AI. Many have come across similar terms in their professional life.
Perceived Strengths	• **Assessment implies evaluation and checks:** participants find the idea of an assessment reassuring, as it implies AI products are being tested and checked.

Perceived Weaknesses	• **'Impact' doesn't speak to risks**: participants take impact to mean how well an AI program works. While this may be important to assess, it doesn't necessarily mean an impact assessment. Participants don't necessarily trust AI just because it works. • **Many feel that this term implies AI creators are also the assessors**: participants assume that those who have created AI will be doing product testing, something akin to impact assessment.
Key Takeaway	While participants think impact assessments are intuitive and can ensure that AI programs are working effectively, they are unsure whether they can provide protection against risk, given what they are evaluating and who they assume to be carrying out the evaluation.
Evidence	*"Impact assessments is positive, because if the AI isn't working, [the impact assessment] will tell you."* *Female, Distrusting Data Sceptic*

Use Case: Personalised Health

Perceived risk: Medium

How this use case affects views towards assurance terms: Participants initially focus on whether it is right to introduce AI to something they feel is far better suited to human involvement. They question whether AI is replacing doctors, or assisting them, showing a much stronger preference for the latter. This use case feels futuristic and, to some, dystopian, making it harder for them to address in practical terms and within the context of assurance. When focusing on assurance, participants especially focus on who would be providing the checks and evaluation. They show a preference for well-known or trusted bodies, such as the British Medical Association (BMA)n or the NHS itself.

Regulatory Compliance

Participants find the idea of AI being compliant with laws and regulation reassuring. In some cases, they call for specific regulation for AI. However, many are doubtful that effective regulatory compliance can be put in place for an industry that they see as being hard to control. Information about this term should directly address how regulation is able to control this industry in the UK.

Term tested: Legally compliant

Level of assurance	**High**
Understanding	**Language is formal but clear:** participants respond well to language that feels official but is still clear for them to understand and conveys its meaning effectively.
Perceived Strengths	• **Participants like to hear laws are in place:** participants react well to the idea that AI will be checked within a legal framework.
Perceived Weaknesses	• **Legal framework has limitations in perceived effect:** participants worry that AI creators will be able to find work arounds or stay one step ahead of any legal framework in place. • **Uncertainty over the laws around AI:** some participants question the status of AI laws in the UK, claiming that many are still being written.
Key Takeaway	While there is a concern about how the law can be applied to AI, this term effectively conveys that an AI service or product sits within a legal framework, a check that most find reassuring.
Evidence	*"Legal compliance means you know they are doing the right thing, and we have the back-up of the law if things go wrong."* *Female, Less Confident Digital User*

Term tested: Compliance audit	
Level of assurance	**Low**
Understanding:	**The term is hard to understand:** most feel that it is vague and question what it means or entails.
Perceived Strengths	• **Audit is perceived to mean thorough, regular checks:** as above (see 'Bias Audit'). • **Compliance has legal connotations for some:** for a minority of participants, the term, in

	particular 'compliance' provides reassurance that AI is being checked within a legal or regulation framework.
Perceived Weaknesses	• **It feels like jargon**: participants feel this term comes across as jargon and lacks substance, contrasting a term like 'legally compliant' that sounds official but is also clear.
Key Takeaway	This term does not reassure participants, as most do not have a clear understanding of its meaning and feel that it comes across as vague and overly technical.
Evidence	*"It's a bit less clear because the public don't actually know what a compliance audit is or what it entails…it just sounds vague."* *Female, Less Confident Digital User*
Term tested: Audit	
Level of assurance	**Medium**
Understanding	**The term is clear:** most understand what an audit is but not all are sure exactly how it can be used to assure AI without more of an explanation.
Perceived Strengths	• **'Audit' means thorough, regular checks;** as above (see 'Bias Audit').
Perceived Weaknesses	• **It lacks detail:** audit doesn't refer to how the checks will be carried out, nor does it imply who will be carrying out the checks. • **'Audit' feels like jargon for some:** although clearer to some than 'compliance audit', others state that 'audit' also feels like jargon, especially within the context of AI.
Key Takeaway	While audit can be a useful term, because of the nature of the check it denotes, it needs to be included alongside a clear explanation or detail setting out what is being audited and who is carrying it out.

Evidence	*"An audit would give me some confidence but I'm still worried about who holds your information."* *Male, Enthusiastic Tech Pro*

Term tested: Conformity assessment	
Level of assurance	**Low**
Understanding	**It is hard to understand:** many feel this term is unclear, feels like jargon and won't make sense to the average person.
Perceived Strengths	• **Assessment is popular:** participants like the idea that AI is being tested and assessed, it shows them that it is being checked and verified before being approved for public use.
Perceived Weaknesses	• **Questions around what AI is conforming to:** based on the perception that a legal framework or regulations are not in place, participants asked how a conformity assessment can be carried out if it is not clear what AI should be conforming to.
Key Takeaway	Although the testing implied by assessment is popular, the use of conformity raises more questions than it answers. Many feel this term feels overly technical without having a clear meaning.
Evidence	*"What is it conforming to? What rules and regulations?"* *Female, Distrusting Data Sceptic*

Use Case: Facial Recognition

Perceived risk: Low

How this use case affects views towards assurance terms: Participants are much more at ease with AI that is currently in use and is part of a technology offering from well-known brands, for example Apple using it in their phones. They assume that this technology has been tested and confirmed

as safe or suitable to use, if it is being included in mass technology. For some participants, a high level of detail, especially when it comes to 'compliance' or 'conformity', feels too technical for them and the average person to understand. They prefer instead to assume that such technology is compliant rather than needing detailed information about compliance to make this assessment themselves.

Certification as a means of demonstrating AI assurance

Key insights
1. **Participants feel positive about certification as a means of demonstrating AI assurance.** It quickly tells them that an AI product or service has been checked and verified for its safety and security. 2. **The organisation that is issuing certification is more important than the type of certification (e.g. product, professional).** Participants want to know that the organisation is independent (i.e. not motivated by profit from developing AI) and has the required expertise to assess AI. 3. **The context in which participants encounter certification influences the extent to which it feels reassuring. As with other areas of assurance, if the use case feels particularly high risk, then certification itself is not sufficient to build trust.**

Views towards AI certification as a means of assuring trustworthiness

Overall, participants feel positive about AI certification as a means of assuring AI. They feel certification is a useful concept that can quickly tell them that something has been assessed and approved. Many participants draw comparisons with certification in other sectors that they are familiar with, such as the FCA watermark on financial products and the British Lion mark on eggs.

As with other products and services more generally, participants are less interested in the process of certification. For example, whilst they know that financial firms are regulated, they are not sure *how* the FCA evaluates firms. In this context, a watermark or stamp of approval is seen as sufficient in most cases.

However, the context in which participants encounter certification is crucial. The perceived level of risk an AI use case poses coupled with who is using the technology (e.g. a teacher, a doctor) is more influential than whether or not it has been certified. For example, many of the less digitally confident participants would not trust a self-driving vehicle, even if it had been certified. Some participants would not trust a teacher using a certified AI product as they feel they are too time-poor to do their own due diligence.

As with AI assurance more generally, the certification body of AI assurance is more important to participants than the process of becoming certified. Participants want to know who the body is, that it is independent and has the required level of expertise. Whilst some feel the Government has a clear role to

play in certification as a means for assuring AI, others question its competence to do so.

Finally, participants feel AI certification can be a way to help normalise AI technology in society. They feel continued use of certification as a means to demonstrate AI assurance can build awareness and trust in products over time, and certification can be a marker to influence which products they choose and feel safe to use.

Benefits of certification as a means of assuring AI:	**Drawbacks of certification as a means of assuring AI:**
• A quick way to tell the public that AI products and services have been reviewed and approved. It therefore takes the responsibility to conduct further research off the end user. • An opportunity to build trust and set a standard for products and services to meet.	• Catching up with the pace of change. Participants question whether certification standards will keep up with constantly evolving innovations. • Time taken to introduce certification means there is greater risk in the short term.

Understanding of certification terms

Participants were shown different terms – 'certified', 'certification scheme', 'accreditation', and 'trusted third party' – to assess their understanding in the context of certification of AI assurance.

Whilst participants claim they are familiar with all terms, the majority of participants show greatest understanding of the terms **'certified' and 'certification scheme'**. Other terms ('accreditation', and 'trusted third party') are less clearly understood.

Term tested: Certified	
Level of understanding	**High**
Understanding	The public find this term clear and easy to understand: they believe it means that an AI system or product has been checked, met certain criteria and been approved by an official body.
Key Takeaway	This term successfully communicates a process of testing of risks by an official body. It can be used to let the public know an AI product or service has undergone certification.

Evidence	*"It's passed certification and gone through various checks and balances."* *Male, Distrusting Data Sceptic*
Term tested: Certification Scheme	
Level of understanding	**High**
Understanding:	Consistent with 'Certification', participants feel this term is clear: most interpret it to mean the types of checks something must go through to be 'certified'. However, participants do struggle to clearly differentiate the definition of 'certification scheme' from 'certified'.
Key Takeaway	'Certification Scheme' can be used alongside 'certified' to talk about how AI is being checked. Participants are able to place it next to other formal processes, checks and approaches, for example 'British Standards'.
Evidence	*"I would look at it in terms of a group that would certify themselves together."* *Male, Less Confident Digital User*
Term tested: Accreditation	
Level of understanding	**Medium**
Understanding	The handful of participants who claim to understand this term define it very similarly to 'certified'. They feel it is when a product or service has met specific standards. However, other participants feel the term is vague and jargonistic. They want to know more detail to understand how it relates to AI.
Key Takeaway	Overall, participants struggle to define 'accreditation' (and how it differs from the other terms). Greater explanation will be required to ensure it is well understood.
Evidence	*"It means you adhere to certain standards."* *Male, Less Confident Digital User*

Term tested: Trusted Third Party	
Level of assurance	**Low**
Understanding	Although the terminology is familiar, participants are unsure whether this refers to an independent body or a business.
Key Takeaway	'Trusted third party' is insufficient to provide reassurance without more detail on who is trusted and how they can receive this verification. Participants question the independence implied by this term and want clarity on the kind of body 'third party' stands for.
Evidence	*"It feels external, but sounds a bit wishy-washy."* *Male, More Confident Digital User*

Certification types

The most important factor when discussing types of certification is that an independent body has carried out the evaluation. Participants are less clear on what the differences between a 'certified product' vs. a product developed by a 'certified professional' will mean in practice. In all of the below, participants want to know *who* has carried out the certification and to be reassured about their competence and motives.

'A certified product' feels the clearest to most participants. They can assume an AI product or service that has been 'certified' has undergone an official check and verification, which feels reassuring to an extent. However, some question who is certifying the product and want to hear more detail about this. In particular, they are looking for confirmation that those involved have both the required level of expertise and independence to perform a thorough and trustworthy check.

'An AI product developed by a certified AI professional' feels less clear to participants overall. A handful of participants who are familiar with chartership in other professions (e.g. engineering) feel this type of certification is clear and reassuring. It lets them know AI products and services have been developed by an expert. However, for others who are less familiar with certified professions, this type of certification raises concerns. They question the motives of individual professionals and want to see reference to 'independence' in this context.

> *"It's a bit reassuring because you think they have a bit more knowledge, but it's still not talking about a separate body."*

Female, Distrusting Data Sceptic

'An AI product developed by a certified AI company' is the least well-received certification type tested. Participants assume it will be the developers of AI who will be carrying out the checks. This is perceived negatively as they assume companies will prioritise profit over thorough safety and security checks. The fact that a company would be certified in this context does not address this concern.

A minority feel that this type might be better for preventing mistakes than certification by a 'professional', simply because they assume a team of experts would be working on a product, rather than an individual.

"'Company' suggests that more people are involved, so it's more reassuring."

Female, Less Confident Digital User