Britainthinks

Insight & Strategy

CDEI | AI Governance

Full report

April 2022

Contents

1 Introduction & key findings

'To support the development of the AI White Paper, it was important for CDEI to engage with the public to understand their expectations for AI governance. The focus was on expectations for three key principles: transparency, fairness and accountability. Engagement was done both qualitatively (via focus groups) and quantitatively (via polling). We covered:

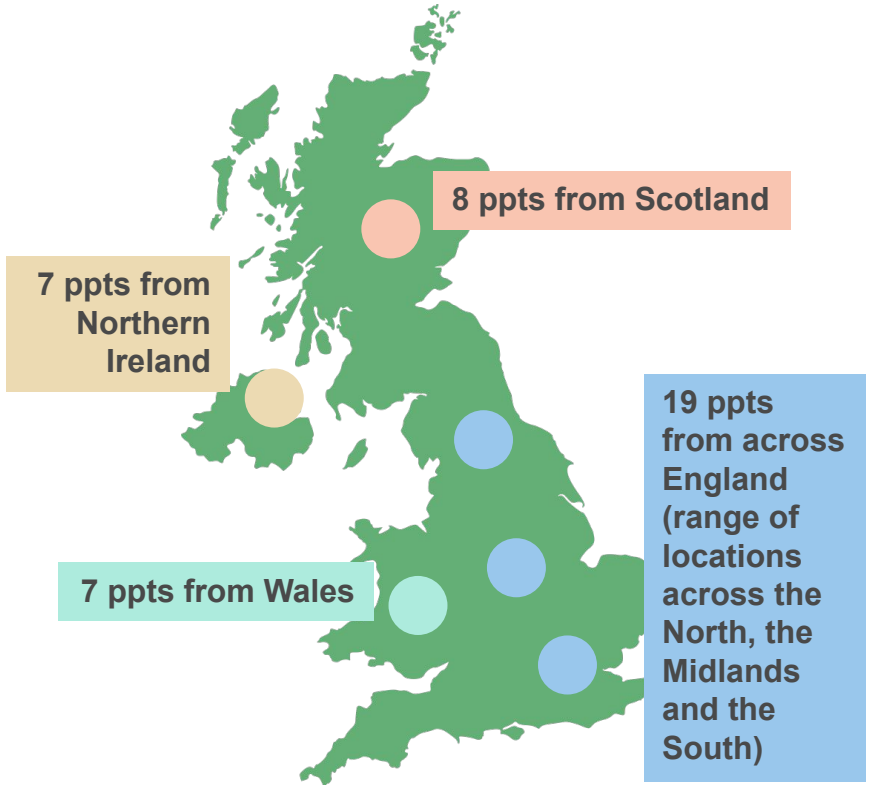
1.	The public's key **hopes** and **concerns** about **AI** and **AI governance**
2.	The public's expectations for transparency, fairness, and accountability in relation to AI
3.	What the public **expect** the government to be doing to **keep them safe** in regards to AI governance

Due to low awareness and complexity of the topic, a three-stage methodology was used to allow time for reflection

Each of the three stages were held a week apart in order to allow participants to share their opinions as their own thoughts developed.

Phase 1	Phase 2	Phase 3
7 online qualitative focus groups	**1 week of online activities completed on an individual basis**	**7 online qualitative focus groups**
An initial qualitative stage to explore **spontaneous perceptions** - including concerns and benefits - of AI, as well as **introducing the three key principles** to participants in both non-AI and AI settings.	**Gathering specific expectations** of governance with relation to the three key principles in the five different use cases	A final qualitative stage to explore the **'why'** behind expectations expressed in the online activities, and to further understand **what was important** to the public in the governing of AI in future
w/c 14th March '22	*w/c 21st March '22*	*w/c 28th March '22*

41 members of the public took part, and were recruited from a range of locations across the UK, as well as:



A mix of…

- Age, gender, ethnicity, socioeconomic group, education and urban/rural living
- Optimism towards technology in the future
- Digital literacy
- Health, including those with physical disabilities, LTHCs, mild/moderate mental health conditions and neurodiverse conditions

A mix of prior experience with some of the use cases for AI…

Advertising and news	3 x participants who had a Google account and who use social media for at least 15 minutes every day
Health	3 x participants who use an app to track an element of their health at least 3 times per week
Financial sector	3 x participants who are in receipt of at least one benefit and 3 who use online banking at least twice a month
Recruitment	3 x participants who had changed jobs within the last three months

Focus group research was supported with a large nationally representative survey with 4,120 UK adults

Who?	**4,120 UK adults** from across all regions of the UK and weighted to be nationally representative.
What?	Polling across **six use-cases** of AI (with three crossing over with the qualitative research) focusing on the themes of transparency, accountability and fairness.
When?	14th - 18th March 2022 with polling provider DeltaPoll

Key findings

1. The public continue to have limited awareness of AI, with knowledge mainly of low-risk use cases that are already in use, e.g. recommendations or advertising suggestions. Other associations are largely futuristic and based on media images, e.g. robots.

2. The benefits of AI are broadly seen to outweigh the risks, with future society deemed to be more efficient with AI's involvement, however the risks are more front-of-mind with strong concern about societal reliance on AI and where this may leave individuals and their autonomy.

3. Participants' views on governance around transparency and accountability are tied directly to the perceived risk of AI's use in any given context. In contrast fairness, and the data AI should be able to use, is seen as a trade off between relevance/utility of a data source and individual privacy.

4. Low familiarity with more complex AI applications makes it difficult for participants to specify what governance they expect. What they do want is the same principles of transparency, fairness, accountability and privacy to apply, along with context specific limitations.

5. The public therefore expect AI Governance to work with these principles but also to develop ahead of detailed public understanding and expectation.

2 Public perceptions of AI

This research set out - in part - to understand the latest perceptions of the public about AI technology.

In line with previous research conducted by BritainThinks on behalf of CDEI, individuals continue to have a consistently narrow view about AI's potential, and continue to be broadly unable to spontaneously identify AI in technology around them.

The public find it difficult to spontaneously identify AI in their day-to-day lives, but are aware the latest technology is at work behind the scenes

When asked about how they thought a set of familiar and unfamiliar examples of AI worked*, most participants referred to ideas such as **'cookies', 'tagging'** or '**algorithms',** versus calling out the term 'AI' specifically.

Therefore, individuals' awareness and understanding of AI continues to be fairly limited to commercial and digital applications of AI, with very little awareness of AI being used in public sector settings.

"I think the whole 'tracking your usage' by apps that you might not even be aware of is potentially the the most annoying or intrusive part of these sorts of technology for me."
Participant aged 18 - 24

**To understand the extent to which individuals are aware that AI runs many of the services and apps they use, we showed a range of different examples of AI and asked participants how they thought these worked. Examples included well known applications such as facial recognition, Alexa/Siri, social media newsfeed recommendations and self-driving cars, as well as some less common applications such as a government system allocating council housing*

The public naturally have highest awareness of AI where it already directly affects their lives

Lower awareness → *Higher awareness*

Participants are very unlikely to be aware of AI's role in wider societal systems that have little to no direct contact with individuals, such as the prioritisation of housing. However, there is some awareness of mass surveillance systems and the role of newer technologies within these, but no ability to identify AI as the main driver behind such operations.

Awareness of AI's role in self-driving cars is less obvious to participants but is also largely unsurprising, mainly due to their widespread media coverage and associations with forward-thinking entrepreneurs such as Elon Musk.

Participants speak most confidently about how searching for something online (or even speaking about it) will lead them to be targeted by an associated product or service online. This is by far the use of AI technology which is most front-of-mind for individuals and is also something many participants express unease about.

"I've seen surveillance cameras, but didn't know AI would be involved with something like that"

Participant aged 35 - 44

"With targeted adverts when you look at an item and then 30mins later you see the adverts on google. It basically see's what you've looked at and guess what you'll like to buy next"

Participant aged 45 - 54

By the end of the research process, most agreed that the benefits of AI outweigh the risks when thinking about the future of society

Examples of AI are most strongly felt to bring **convenience** and **efficiency** to both individuals' lives as well as large-scale processes that require preliminary sorting before being addressed by a human. In addition to this, a number of other benefits were raised on prompting, such as**:**

Removing bias by humans	Improving accessibility for all	Creating safer communities	Increasing connectivity
Participants feel that AI has the potential **to help remove bias** (even unconscious bias) from decisions made by humans, e.g. bias towards appearances, ethnic or cultural backgrounds or personal connections.	AI is seen to **make a large range of aspects of daily life significantly easier**, something which is seen as particularly beneficial for people who had struggled beforehand as well as creating new opportunities for people who had previously been excluded from parts of society due to a physical or mental disability.	For those living in urban areas, and in particular women, AI is seen to have great potential **to make communities safer** with the use of cameras which may detect crime. This is in the context of more extreme deprivation and increasing crime rates.	Older participants in particular highlighted the role AI technology has already played in **connecting them with other people in their community by** suggesting certain groups to join, something which has been particularly beneficial during and following Covid-19.

However, risks and concerns are far more front-of-mind, often influenced by examples of AI in media and concern about privacy

Perceived risks and concerns about AI technology are dominated by the potential for AI technology to one day have **more control over human behaviour** than individuals do, and the associated inability to know **when this line has been overstepped.** Three more specific risks, all of which are tied to this concern, were also mentioned:

Invasion of privacy

Participants across all age groups express concern about the lack of boundaries within the data that AI appears to have access to. All participants were at least somewhat aware that devices seem to **listen to their everyday conversations**, evidenced by the sudden appearance of related content. While a handful of participants feel this is a part of life that has to be accepted, most were unhappy with this and felt this went too far in leaving them with no privacy in their life at all.

AI's role in influencing public opinion

There is broad awareness amongst participants that algorithms within AI technology can lead to individuals consistently seeing only one view or opinion. There is also awareness, though to a lesser extent, that AI has been used by hostile actors to **influence individuals' beliefs,** predominantly through social media. Participants show concern about the potential for AI to be used maliciously again in future in order to manipulate important outcomes such as elections or to create echo chambers and division within society as a result.

Negative health outcomes

Participants show concern about society as a whole becoming **too reliant** on the convenience brought by AI technology and becoming more sedentary as a result. Participants show concern about the long-term impact this would likely have on individuals physical and mental health, and the associated costs for the already-struggling NHS further down the line to correct this.

The public are broadly trusting of the decisions made by AI in use cases they are familiar and more comfortable with, however as this familiarity decreases, so does their level of trust

*Decisions made by AI tend to be **more trusted** when:*

- Participants **already have experience** of AI technology being used in this scenario and where they have been able to build trust in the AI's decision-making abilities;
- There is **low potential impact** on individuals if decisions made by the AI are incorrect (Alexa/Siri, Face ID, search engines recommending articles);
- The **potential for human bias is high** and where the use of AI technology therefore is seen to largely remove this;
- There is a great need **for safety and precision** and where the **risk for human error is very high** (e.g. medical environments).

"It [AI] speeds things up. Like with [google] maps, it's a lot quicker to just put in your phone where you need to go than to look, you know, on a paper map or anything, and it's always accurate"

Participant aged 18-24

*Decisions made by AI tend to be **less trusted** when:*

- The risk of an incorrect decision made by AI has a **high negative potential impact** on people (e.g. Government using automated systems that decide prioritisation for council housing)
- There is **no back-up system in place** if the AI technology fails altogether (Self-driving cars)
- **A more nuanced decision needs to be made** that encapsulates an individual's emotions or life circumstances (e.g. the impact of the pandemic on finances, mental wellbeing)

"I don't AI can make decisions when people's lives have been affected. It's hard to explain when things have happened in your life to a chat bot. You need a person for that"

Participant aged 45 - 54

3 Real world applications of AI

Qualitative testing of five use cases against the three key principles

Eight use cases were tested across the qualitative and quantitative strands of this research.

The following section takes an in-depth look at how the public felt AI should be governed within each of the 5 use cases which were qualitatively tested against three key principles as defined by the Office for AI:

Transparency	*In each situation where a decision has been made using AI, **what information would you as an individual expect to be available**? Where would you expect to find this?*
Fairness	*To what extent is it fair for decisions using AI to be based on **specific characteristics** (e.g. age, gender, ethnicity)?*
Accountability	*Where decisions made using AI are not agreed with, how should the affected individual be able to **exercise their rights**?*

The eight use cases tested were as follows:

1. Recruitment	Quant	Qual	A supermarket recruiting for a role uses AI to assess applicants' answers and determine who is invited for a follow-up interview
2. Mental Health Chatbot	Quant	Qual	A healthcare organisation uses a chatbot to support people with their mental health needs, and puts users in touch with MH professionals
3. HMRC Tax Fraud	Quant	Qual	AI uses tax records and other information to decide if an individual's behaviour could be fraudulent and flags cases to investigators
4. Music Streaming	Quant only		Using demographic information and previous listening habits, AI recommends music that users might be interested in
5. NHS Transplant list	Quant only		AI uses information about patients requiring an organ transplant to decide where they should be placed on the organ transplant list.
6.Police Facial Recognition	Quant only		AI processes live CCTV and compares images of people's faces in public to those on a police 'watchlist', alerting the police and deploying officers if a match is found.
7. Shopping suggestion	Qual only		Using purchase history and demographic information, a well-known supermarket website uses AI to suggest potential purchases for you
8. News recommendations	Qual only		Your web browser suggests news articles based on previous articles you viewed, your demographic information and browser history

3.1 AI in recruitment processes

Use case: A large supermarket is selecting potential candidates for a job on the shop floor and uses technology to select which candidates they will invite for a final interview. As part of the application, candidates are asked to submit answers to several questions via video. The technology then assesses applicants' answers and scores them based on several criteria with no human involvement. Those candidates with the highest scores are automatically invited to interview and the other candidates are rejected.

In this use case, most important to the public is having a clear explanation of the criteria used by the AI to make decisions

In order to understand how they will be judged in an AI recruitment situation the public told us they expect:

1. Companies to clearly signal to applicants that their video application will be **judged by a computer system/AI** and not by humans.
2. Companies need to share clear upfront information at application stage about **the criteria being used by these AI systems** to decide who will progress to the next round.
    a. In particular, the public are doubtful that AI could read and interpret subtle cues like personality, tone of voice, emotion or emphasis, all of which were felt to be important elements of communication. They wanted to understand how (if at all) these factors would be considered.

"I'd want to know how the AI is programmed to make those decisions? How is it deciding who goes through and who doesn't?"

Participant, 55-64

"If everyone knows what the criteria it is, then the process seems very fair. There's no room for nepotism or anything like that."

Participant, 65+

Aligning AI governance with current employment regulations on discrimination will be key to ensuring processes are seen as fair

Trait	Fair*	Unfair*	Rationale
Name	16	25	Not seen as relevant information for the AI's decision-making, and even those who class name as 'fair' information to use still expect the AI to have no bias towards different names.
Gender	13	28	Not seen as relevant information for the AI's decision-making based on the view that gender has no impact on someone's ability to carry out a job role.
Age	16	25	In light of laws against age discrimination, age is seen as an unfair characteristic to include, with a few exceptions like roles that include selling alcohol.
Ethnicity	8	33	Ethnicity is seen as having no bearing on an applicant's ability to carry out this job, and current discrimination regulation is seen as another reason why this is is unfair information to use.
Accent	8	33	Along with the fact that accent is seen to have no impact on someone's ability to carry out this role, many participants question the AI technologies' ability to interpret all different UK accents.
Skills and experience	38	3	This information is seen as undoubtedly the most important for AI to pick up on in order to make decisions about who should be put through to the next round and who shouldn't.
Tone of voice and mannerisms	26	15	In this role which likely includes an element of customer service and engaging with vulnerable audiences, most participants feel it is important to take into account an applicant's personality.
Interests and hobbies	25	16	While some feel this information is irrelevant and therefore unfair to take into account, most feel this could help to determine an applicant's overall fit for the role, including transferable skills.

*Fair and Unfair columns count the number of participants that scored each characteristic as either option on the online community.

Being able to challenge a decision or obtain feedback about why an application was not successful is important to participants

Expectations for accountability

In the context of applicants being *un*successful, participants have inherently **lower trust** in the AI's decision-making compared to decisions made about their application by a human. As a result, and given the potential importance (financially) of being successful in the recruitment process, it is even more important for participants to be able to **access feedback about why** they have not progressed further in this scenario, or to have the option to **challenge the decision entirely**.

"It's very important to have feedback, and to have the ability to challenge the response overall."

Participant, 45-54

How this needs to happen

Participants have modest expectations for *how* this needs to happen, and feel that being able to **reach out to the company** they were applying to would suffice. That being said, even though these decisions have been made by AI, participants still expect to receive a **detailed and personalised response** as to *why* they were not put through to the next round.

"It needs to be a proper justification with detailed feedback, not just a generic response."

Participant, 25-34

3.2 AI in a health-focussed chatbot

Use case: A healthcare organisation uses a chatbot on an app or website to respond to individuals and support them with their mental health needs. A chatbot is a type of technology that has text conversations with a user as though they were interacting with a human. The technology automatically processes a user's messages and tailor responses to the user's needs based on this. In this instance the mental health chatbot tracks users' mood, offers advice and emotional support, and puts users in contact with mental health professionals if they are needed.

Due to the vulnerability of the users of this service, participants have high expectations for the information that will be provided

However, participants' **incredibly low trust** in AI technologies' ability to recognise emotion in the same way as a human would means that, in this scenario, participants are looking for more information to be provided about the **design and operation of the service** as opposed to information specifically about how decisions are being made.

Firstly, participants feel it must be made clear upfront that **you are speaking with a bot** as well as **who this service is for,** i.e. not for individuals who need urgent support in a crisis, and that these individuals should instead do 'X'.

Participants overwhelmingly express their need to know that there is the **ability to connect to a human** – even immediately if necessary, as well as the AI providing relevant signposting information so that users are informed about other potential sources of support (acting almost as a triage service).

Participants also feel it would help to build trust in the service by understanding **who was involved in its' design** – for example publishing that this service was created in collaboration with mental health professionals.

Similarly, to increase trust in the ability of the AI to provide support, participants would like to know about the extent to which a **system has been tested** prior to being put into use.

"I would need to know that I am speaking to a bot, and that I can talk to a human at any point in the conversation."
Participant, 55-64

"I'd want to know that the chatbot had been designed with professionals within that field and tested well enough before being released to the public"
Participant, 35-44

In this use case, far more information is seen as fair and important to use, including personal identity information

Trait	Fair*	Unfair*	Rationale
Name	**23**	**18**	AI using a users' name could provide better support and comfort to a user as well as being important for referrals, however others feel this is irrelevant and unimportant data for the AI.
Gender	**30**	**11**	Most feel this is fair to use on the basis it may help the AI to understand people's experiences of mental health.
Age	**32**	**8**	Seen as relevant and fair information to use in order to help build a picture about the user.
Ethnicity	**19**	**22**	Those who agree this is fair data to use generally feel it may help the AI to provide better support by having a better understanding of a users' background, however others feel this is irrelevant and therefore unfair data to use which may also make a user feel uncomfortable.
Location	**31**	**10**	Most feel it is fair to use this information in case the user requires emergency support with whom location data might be shared *or* so that the AI can advise on local support services for the user.
Interest and hobbies	**29**	**12**	Fair since this information helps the AI to build its' understanding of the user as a person as well being able to suggest potential strategies or coping mechanisms to improve their condition.
Mental health record	**37**	**4**	Participants feel the AI is likely to make more appropriate suggestions knowing about the users' mental health history, as well as potentially helping to overcome unnecessary initial dialogue.
Precise words and tone used in the messages	**33**	**8**	Important information for understanding the user and their state of mind, including the urgency of a user's situation, with others doubting the AI's ability to interpret these characteristics.

*Fair and Unfair columns count the number of participants that scored each characteristic as either option on the online community.

Unsurprisingly, the importance of speaking to a human in this use case is very high, as well as the need to collect feedback

Influenced by the perceived infancy of AI's use in scenarios such as this, participants show great concern about the potential for AI to provide **poor quality support** to vulnerable users or to even provide **incorrect information** which may lead to high-risk negative consequences for the physical and mental health of individuals using the service.

While a support session is live		*Once a support session has concluded*	
Participants believe the chatbot should have the capability to **recognise when it is not able to provide adequate** support to a user and instead refer a user them a human.	Equally, participants show concern for users who may be given **misleading or incorrect information** by the AI and who may need to access support from a human.	Including a function for **users to provide feedback** on their session is seen as a way for the technology and developers to consistently have access to the information needed to improve and build on the service.	Participants also suggest that the AI technology in this scenario should undergo **regular reviews** in order to ensure it does not go for long periods of time under-performing.

"If it [the AI] cannot provide the individual with a suitable answer, then it should automatically be transferred to a human."
Participant, 55-64

"There should be an option to provide feedback on the quality of support received in order to help developers understand how to improve the service."
Participant, 25-34

3.3 AI in HMRC Tax Fraud Investigations

Use case: Imagine a scenario where HMRC (the government department responsible for taxes) uses technology to identify people who are likely to have committed tax fraud. The technology uses tax records and other information to decide if an individual's behaviour could be fraudulent and automatically flags these to investigators

In this use case, the public want to understand what information is used, outside of tax records, in order to flag someone's profile

In relation to transparency, there are **two pieces of information** that participants would want to see should they be contacted by HMRC as part of an investigation:

Participants want to be presented with a **report** which outlines the **specific piece of information** that caused them to be flagged, as well as an overview of the **different types information** - in addition to their tax records - that are used by the AI technology to identify fraud.

Participants would like to have a clear explanation of how an individuals' (their) personal information was obtained **in line with current GDPR regulation**.

"I would like to know the criteria used that caused me to be flagged up, so that I can make sure everything could be cleared up and clear my name." Participant, 65+

"I want to know how HMRC got to their decision, what information they used and how this data was obtained and whether it was got in line with the GDPR rules." Participant, 45 - 54

A certain amount of personal information is seen as necessary in order to assist the AI's screening of fraudulent activity

Trait	Fair*	Unfair*	Rationale
Name	28	12	Using an individual's name is seen as fair by most since the investigation may require this in order to ensure the correct individual is being identified against the tax records, whereas others feel this is irrelevant data.
Age	26	14	Seen as broadly fair information to use since different age groups may be subject to different tax regulation and in order to assist with personal identification in the case that investigations are taken further.
Ethnicity	11	29	Using this information raises red flags for participants around discrimination and potential racism if taken into account by the AI, with all agreeing that the AI should show no bias based on an individuals' ethnicity.
Gender	16	24	Most feel it is fair to use this information in order to help confirm the identity of the individual being investigation, however others feel it is unfair to use this information since gender plays no role in determining someone's likelihood to commit fraud.
Tax records	38	2	This is the main piece of information used by the AI to determine fraudulent activity, and participants are therefore consistent in their view it is fair to use this data for the initial screening as well as the investigation.
Social media posts	15	25	Using this data is seen as a potential invasion of privacy for most participants, however even those that feel this is unfair to do acknowledge that it may help to identify those who are not abiding by tax regulation.
Location data	24	16	Most feel this information may help the AI to understand an individuals' behaviour with more accuracy and as a result, their associated risk for fraud, however a significant proportion feel this information breaches hat they are comfortable with in terms of their privacy.

*Fair and Unfair columns count the number of participants that scored each characteristic as either option on the online community.

Being able to challenge the AI's decision in this use case is most important due to its' potential to negatively impact someone's future

Participants make the assumption that incorrect flags on an individual by the AI will be corrected in the next stage of the process when a flag is further investigated by a human.

However even with this in mind, participants exhibit low trust in both the AI's and investigator's ability to spot potential errors from within the AI's decision-making.

As a result, following being contacted by HMRC, participants have a **strong expectation to be able to speak to a human at HMRC** in order to discuss this suggestion of fraud, especially given the high potential impact of such a scenario on someone's future.

"A person may have a reasonable excuse as to why they have suddenly come into money or splurged on a few rather expensive items. A human would be able to reason why certain anomalies have occurred. Again, speaking to the investigator would be a chance to check whether the AI programming is correct and allow a chance to check any faults. ."

Participant, 35-44

"Being able to speak to someone is very important, especially when clues have been picked up from social media. All may be not as it seems thinking about the perspective of the AI."

Participant, 25-34

3.4 AI in shopping suggestions

Use case: You create an online account with a well-known supermarket website, and add a number of different items to your basket to buy. Before you complete your order, the website makes several recommendations for additional items to purchase based on your browsing history and the items that you currently have in your basket. The platform also promotes items to you based on the purchase history of other people that have similar characteristics to you and based on information you gave about yourself when you created your account (i.e. your age, gender, and address). Finally, the platform also makes recommendations for additional products to buy based on items it predicts you might need in the future, such as clothing for a toddler if you had previously been buying items for a baby.

Participants have lower expectations for transparency in this use case, but do want some control over how their data is managed

Participants are supportive of the use of AI in this use case and feel it already provides suggestions which are broadly helpful and relevant. There are some expectations for understanding more about how these decisions are made, but a greater focus on being able to tweak the suggestions they are shown.

Participants are mostly looking for **a simple explanation of which information of theirs is being used** to form these suggestions, as well as understanding where this was taken from (i.e. specific websites).

To further boost transparency, there is some expectation to have a simple explanation of ***how* the AI follows an individuals' behaviour**, for example is this done using cookies or something else?

Participants are also set on having the ability to **entirely opt-in / opt-out** of particular types of information being used, as well as having the ability to **tweak specific pieces of data** being used, e.g. if someone purchased a one-off gift for someone else, they may not want this to be taken into consideration in forming their own shopping suggestions.

"My main concerns are around consent and people having the ability to opt in and out of those things, and being open and transparent on what the website is watching."

Participant, 25-34

"It is supposed to be very easy to opt-out, but now it's much more difficult and deliberately made obstructive. In a lot of cases, it's much easier to just say yes accept all."

Participant, 35-44

Data taken from outside of an individual's interactions with this supermarket or the sector more broadly are seen as unfair to use

Trait	Fair*	Unfair*	Rationale
Name	21	19	Around half state this information is unfair to use since someone's name would not have an impact on their purchasing behaviour, with others only wanting this data to be used to personalise their shopping experience.
Age & gender	35 / 29	5 / 11	Most feel both age and gender is fair information to use so as to help the AI form more relevant suggestions based on different groups, as well as to avoid suggesting age-restricted products to those under a certain age.
Ethnicity	21	19	Participants from ethnic minority backgrounds are more likely to feel this is fair information to use since it may help suggestions to be tailored to their lifestyles, whereas other participants feel this is irrelevant information.
Purchasing + browsing history from the same website	40	0	Participants expect this information to be used, and feel it is the most relevant – and as a result fair - information for the AI to be using in order to form suggestions that are being shown on the same website.
Browsing history from other websites	17	23	Participants feel this is too intrusive and express their concern for having no privacy at all in their online activity, however those who feel it is fair see it as another way to increase the potential relevance of their suggestions.
Purchase history of those who added the same items to their basket as you	21	19	Around half feel this information would help them to see more relevant suggestions, particularly when trying out new recipes for example, however other participants do not feel that using this information would allow for individuals' own tastes and preferences.
Location data	12	28	Most question how the supermarket would have access to this information and feel uncomfortable with this information being used, adding that they feel this would be irrelevant in forming relevant product suggestions.
Social media posts and interactions	12	28	Most feel that using this information would be too intrusive of their privacy and see it as particularly irrelevant in the context of a supermarket shopping experience, whereas others feel this information could be helpful to use.

*Fair and Unfair columns count the number of participants that scored each characteristic as either option on the online community.

Britainthinks
Insight & Strategy

Participants have no expectations to speak with a human about a suggestion, but want to be able to amend these easily themselves

Participants expect that the two ways AI may go wrong in this use case are that it may show an **inappropriate product suggestion** (e.g. showing baby products to someone who is struggling to conceive) or that the suggestions may be **repetitive** over a period of time and therefore have little relevance. Compared with the other use cases tested, these two outcomes are seen as relatively **low risk** and **low impact**.

As a result, participants' only expectations are that they're able to easily mark a product as '**not to be shown again**' as part of a system that works **effectively** and which will maintain these chosen preferences as time goes on.

"It is only making recommendations. It is not making life decisions or decisions around ethical / moral issues or those which relate to a person's health or wellbeing. I wouldn't need to speak to a human in this instance." Participant, 45-54

"I think if it was ruining your shopping experience because it's so repetitive, you would want to challenge that. Why is this happening?" Participant, 55-64

3.5 AI in news article recommendations

Imagine a scenario where you open your web browser and log into your profile. This could be something like Google Chrome, Firefox or Safari. On your home screen, you are shown a range of news articles that the technology has selected for you based on a range of factors described below. The articles you are shown are stories that are similar to those you have read before, such as stories from the same source or about a similar topic; stories that other people who are similar to you have clicked on based on information you gave when you created your account, including people of a similar age and location or stories that the technology has predicted you may read based on your wider browsing history online (e.g. other websites you have been to recently).

Participants have few further expectations in this use case outside of understanding more about how these news suggestions are formed

Compared to other use cases tested, the public's expectations for transparency in this scenario are **relatively low**. While the data used by the AI to suggest article recommendations is deemed to be equally as personal as in other scenarios, the potential negative impacts on the individual of poorly-made decisions are not seen to be as great.

Participants' desires broadly focus around understanding more about **how** article recommendations are formed**.**

Participants are mostly looking for **a simple explanation of which information of theirs is being used** to form these recommendations, as well as understanding where this was taken from (i.e. specific websites).

"I'd want to know how it got there, i.e. what data was used to create this recommendation?."

Participant, 25-34

Several participants are looking for something simpler, such as information which states **'You were shown X because you looked at Y'**.

"I think for me, something like 'we sent you X because you showed an interest in Y' would be enough."

Participant, 35-44

Creating suggestions from social media engagement or real conversations is seen as unfair, unlike using recent search history

Trait	Fair*	Unfair*	Rationale
Name	13	27	Most agree that someone's name would have no impact on the types of articles recommended to them and so class it as unfair information to use.
Age & gender	27 / 23	13 / 17	Age and gender are both seen as helpful and fair pieces of information to use in order to to better tailor news article recommendations. Age is seen as particularly important to include since it would allow the AI to recommend age-appropriate articles to individuals who are much younger.
Ethnicity	17	23	Participants are divided on whether this is fair data to use, with some feeling this would help to show news that was relevant to their cultural and ethnic background, whereas others feel this data should be off limits.
Search history from your browser	34	6	Generally, the search history from your browser is viewed as the most important way for the AI to understand what might form relevant news recommendations for someone, with only a handful of participants disagreeing.
Location data & recent purchase history	22 / 23	18 / 17	Around half feel it is unfair to use these pieces of data due to a view that they are irrelevant in forming news recommendations. However, the other half feel these pieces of data are fair to use: location data so that you can be shown local news; purchase history to see news and information about products you have bought.
Recent conversations	10	30	Most participants are highly uncomfortable with this data being used feeling this would invade their privacy and leave them questioning what in their life would be left as private if this data were to be collected.
Social media posts and interactions	13	27	Similarly, most also do not want this data to be used, feeling this would be a step too far in invading their privacy with some also questioning how the AI would have access this data in the first place.

*Fair and Unfair columns count the number of participants that scored each characteristic as either option on the online community.

While the public show lower concern about the possible negative impact of AI in this use case, some safeguards are felt to be key

The public do not feel they would want to speak to a human in order to challenge a decision made by AI in this use case. However, what *is* important is having an **amount of control** over what they are seeing and having the **ability to individualise** this based on their preferences.

This is raised as particularly important to consider when it comes to **vulnerable audiences**, e.g. **younger people** who may be more easily influenced by news recommendations, or for individuals who suffer with **mental health conditions** (e.g. anxiety) and who might experience a negative health effect from repeatedly seeing certain topics, e.g. the Ukrainian war.

Several participants describe how some platforms **already have certain features** that allow for a sense of control over what they are shown. For example, YouTube and Google are called out as having functions which ask for **satisfaction ratings** of the recommendations that are made, including having the option to quickly state you **do not want to see** similar topics again. The public would like to see these **quick and simple-to-use functions** in more widespread use across platforms that use AI to make recommendations.

"I do not feel it is important to speak to a human, but you should be able to stop articles that you deem irrelevant to oneself. ."
Participant, 25-34

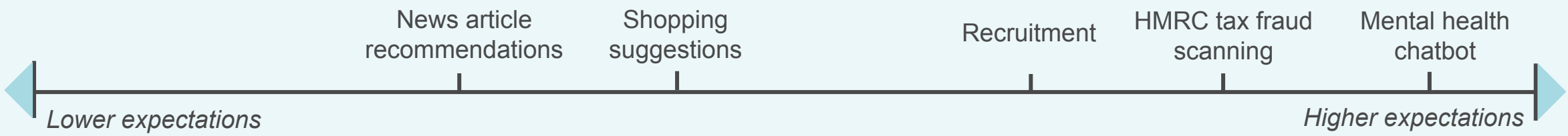
"I don't think it would be very important to me to speak to someone if I could change it in the settings myself."
Participant, 35-44

4 Implications for AI governance

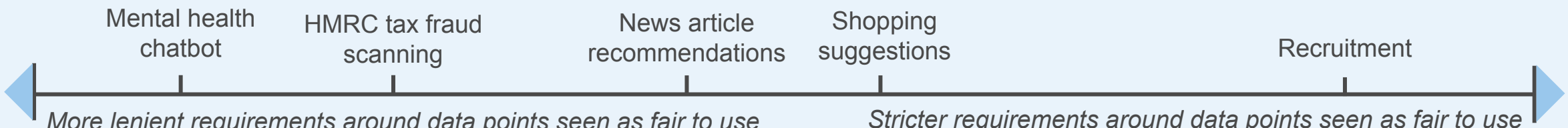
This research has shown that the public have consistent expectations for AI governance based on the uses of AI they are already familiar with, and for use cases they were introduced to during the research.

Use cases that participants are less familiar with tend to elicit lower levels of trust, and therefore expectations for stricter governance

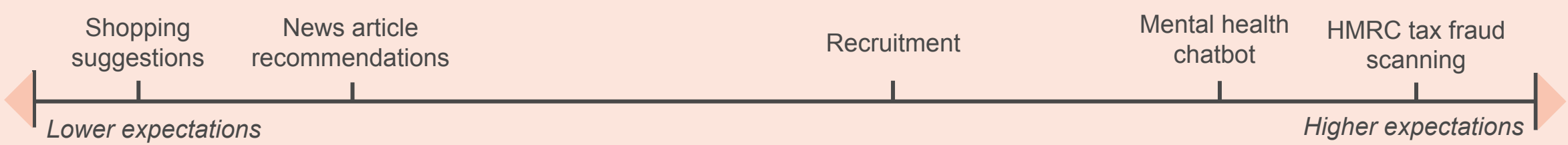
Transparency



Fairness



Accountability



Participants consistently want to know what data of theirs is being used by AI, and where possible, have some ability to control this

Expectations for transparency are highest in use cases with which participants are least familiar and in which they have the least trust as a result (i.e. mental health chatbot/tax fraud).

However, at the very minimum, governance around transparency needs to achieve two things:

1

Ensure that individuals are made aware that AI is being used in decision-making

2

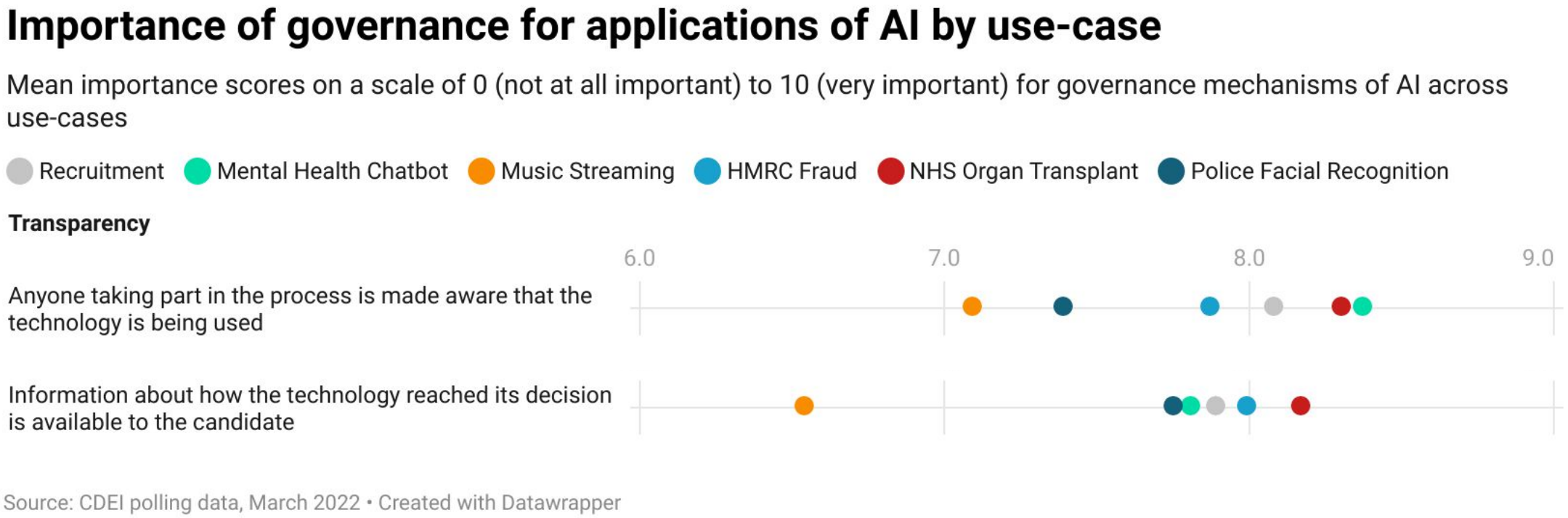
Ensure individuals understand (or can access information which tells them) which data is being used about them to make by the AI system to make decisions

"Knowing that AI is involved in the process is really important. I'd want to know if it wasn't a human making these decisions."
Participant aged 25 - 34

"I want to see the details of what information will be used - what criteria am I being assessed against?."
Participant aged 35 - 44

"I would want to know how my data has been accessed, so where it's be taken from and how it's being used"
Participant aged 45 - 54

Polling also highlights the importance of making people aware when AI is being used



There is an expectation that individuals are made aware when AI is being used, however, the expectation that information is made available about how AI reaches a decision is stronger for use cases which have greater potential to lead to personal harm.

Putting governance on transparency into practice

Mandating useful transparency

- We know that people want to know what data about them is being used and where possible, to have some amount of control over this.

  However currently, opt-in and opt-out pop-ups on websites/apps are seen as **deliberately inaccessible and obstructive in their design.**

- Governance needs not to only mandate transparency of data used by AI and control over this where possible, but also make **more intuitive and accessible processes for controlling this** a key priority. As a result, individuals will feel they have greater control over their own privacy in a context where this is currently perceived to be at risk.

Going beyond data use

- In higher risk use cases where it is deemed more important that the AI makes the correct decision, participants want additional reassurances about the quality or performance of the AI to feel they can trust it.
- They expect governance to ensure more information is provided about how the AI has been ***designed*** (not how it works). This includes sharing information about:
  - The extent to which AI systems have been **tested** before being put into use
  - The extent to which systems were designed alongside **experts** in the field the AI is operating in.

To be seen as a worthwhile use of technology, participants feel an AI application needs to be *less* biased than a human equivalent

It is important to participants that decisions are made '**fairly'.** Upon further exploration, 'fairness' in decision-making according to participants involved three main components**:**

Being free of bias

A 'reasonable' decision – according to participants – is one which is free of bias. That means being based on relevant criteria for the decision, e.g. skills for a recruitment task.

"If it can remove bias, then that has to be seen as a positive thing. "
Participant aged 35 -44

Equal conditions for all

In order for decisions to be made fairly by AI, all individuals affected by a decision **should have equal opportunity and be subject to the same criteria**. This includes ensuring those who have additional needs (such as digitally disengaged people), have their needs factored into the conditions of decision making. As a result…

"You want to make sure that each person is getting the same treatment, whether it's applying for a mortgage or a job, that everyone is treated evenly"
Participant aged 45 - 54

Exclusion of key criteria

Some criteria are felt to be inappropriate in making any kind of fair decision:

- Age
- Ethnicity
- Disability
- Accent

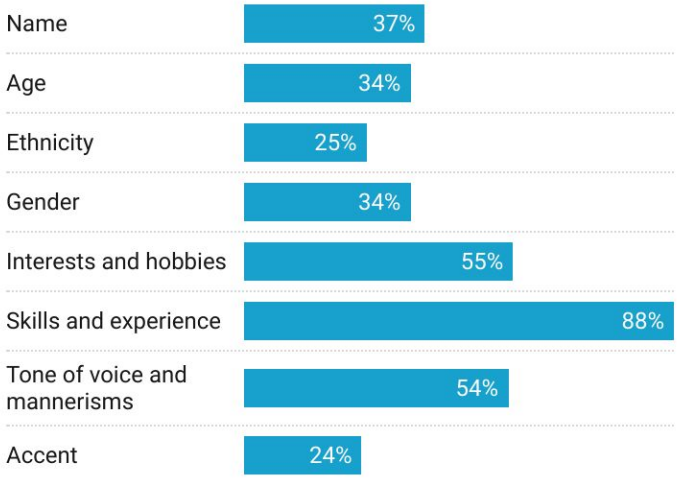
By removing these criteria from decision-making, AI is seen to make at least *as fair* decisions as humans making similar judgements.

"I want to know that the decision being made is being done so on the basis of skill, or aptitude, and not based on where someone is from." *Participant aged 55- 64*

Polling indicated respondents expect only data directly relevant to the decision making to be used by AI to ensure fairness

Recruitment use-case

Percentage of respondents who believe that it is acceptable for technology to use information when making a decisions

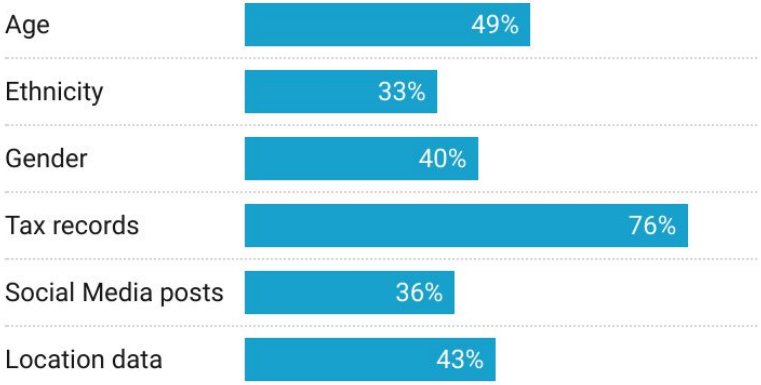


Source: CDEI polling data, March 2022 • Created with Datawrapper

Respondents reported the lowest levels of acceptability for demographic characteristics such as age, gender and ethnicity being used in the AI decision making process.

HMRC fraud detection use-case

Percentage of respondents who believe that it is acceptable for technology to use information when making a decisions

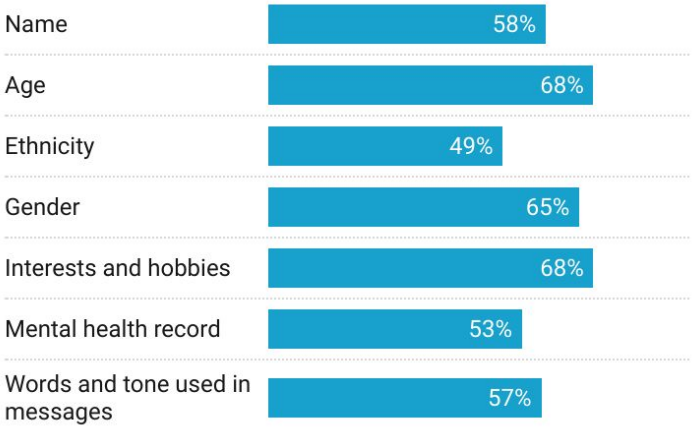


Source: CDEI polling data, March 2022 • Created with Datawrapper

Higher levels of respondents reported that demographic data was acceptable to use in this application compared to recruitment, however, the overall expectation from respondents was that tax records were the most relevant and acceptable to use.

Mental health chatbot use-case

Percentage of respondents who believe that it is acceptable for technology to use information when making a decisions



Source: CDEI polling data, March 2022 • Created with Datawrapper

Many respondents reported that it was acceptable for demographic and wider characteristics to be be used by AI, participants in the focus groups suggested that these factors could be relevant to how the technology should interact with users.

Britainthinks
Insight & Strategy

Putting governance on fairness into practice

Trading off relevance, privacy and bias

- In low-risk use cases such as online news recommendations, people still wanted to restrict the categories of data used to those that were strictly relevant to the purpose of the AI so as to preserve their privacy as far as possible.
- Conversely, for some of the high-risk use cases, people were happy for more personal data to be used if it was seen as essential to the AI making accurate decisions. In these cases, they felt that it was more important for the AI to be effective than to avoid the risk of discrimination.
- Governance will need to balance these three priorities of relevance, privacy and bias.

Respecting boundaries of data use

- **Social media:** Data about people's social media engagement is seen as off-limits since it is perceived as largely unhelpful in forming recommendations, but also because participants are attempting to retain a sense of personal privacy.
- **Listening data:** One of the most consistent and strongest findings was opposition to AI using data from overheard conversations. This was seen as the biggest breach of personal privacy boundaries, even though many suspect it is happening already.

The importance of challenging a decision in different use cases is proportionate to the perceived personal impact of a wrong decision

Lower personal impact ⟷ *Higher personal impact*

News recommendations ***Shopping suggestions***

Challenging AI by tweaking your own preferences

In cases where participants had lower concern about the potential personal impact of a wrong decision, they were more likely to be willing to deal with the 'challenging' of a decision themselves via tweaking the data that was used about them for the future.

"I don't think if AI recommended the wrong product, you would need to speak to someone. You can flag it as inappropriate or make a complaint by email"

Participant aged 18 - 24

Recruitment ***Mental health chatbot*** ***HMRC Tax Fraud***

Requirement for human intervention

In use cases where the potential impact on a human is higher following an incorrect decision by AI, participants express a strong desire to speak to a human to be able to resolve issues. This is most commonplace in use cases when people feel AI lacks understanding/empathy (mental health chat bot) or an incorrect decision causes a person significant distress (HMRC tax fraud detection).

"I'm not sure AI would even have the ability to triage mental health care – so you need to have a button where it puts you through to a person straight away"

Participant aged 35 - 44

Polling data also suggests that human intervention is important for use-cases where negative impacts for individuals are more likely

Importance of governance for applications of AI by use-case

Mean importance scores on a scale of 0 (not at all important) to 10 (very important) for governance mechanisms of AI across use-cases

For the ‘low risk’ use case participants felt it was less important that they would be able to challenge the decision.

For more ‘high risk’ use cases there are strong expectations of being able to challenge decisions with a human if individuals felt the technology had made the wrong decision

Putting governance on accountability into practice

How realistic are the public's expectations?

- Participants had limited understanding of what accountability might look like for AI applications.
- The two options they did identify were:
  - Having the ability to amend which of their data is used by the AI system.
  - Being able to be referred to a human to challenge a decision.
- However as AI use becomes more common these options may become more burdensome for providers and users. Participants struggled to identify alternatives, suggesting a gap for governance to fill.

Protecting vulnerable groups

- Participants are concerned about using newly-created AI systems with groups that are deemed vulnerable, for example those with physical or mental health issues, children or the elderly.
- They expect a higher level of scrutiny and control over the use of the technology with these audiences. This includes ensuring that transparency information is accessible, that consent/control over data is informed, and more generally that an individuals' needs are taken into account.

Perceptions of AI technology and expectations for governance differed by ethnicity highlighting the need for further research

	Recruitment use-case			HMRC use-case			Mental health chatbot use-case		
	Black	Asian	White	Black	Asian	White	Black	Asian	White
% expect positive impact on society	60%	62%	35%	51%	57%	45%	76%	40%	34%
% believe it is acceptable to use ethnicity as an impacting factor on decisions	34%	24%	24%	24%	43%	33%	73%	34%	50%
% think only humans should make this decision	15%	14%	41%	25%	24%	29%	71%	19%	46%
% think technology can be used but the final decision should be made by a human	72%	77%	48%	68%	45%	57%	22%	74%	43%

Results from polling data: n = 3,568 White, 129 Black, 183 Asian (unweighted), other ethnicity groups with n<100 not included in table. Further research is required to investigate findings as sample sizes are small and other impacting factors such as age or region may be impacting results.

A major limitation of engaging the public on AI is their current understanding of the scope and nature of AI systems.

Current associations are largely limited to examples of machine learning and predictive algorithms.

For less familiar use cases or more advanced technical approaches, participants don't have detailed expectations of governance.

Despite this low level of understanding, individuals do make a set of assumptions about AI technology and therefore AI governance

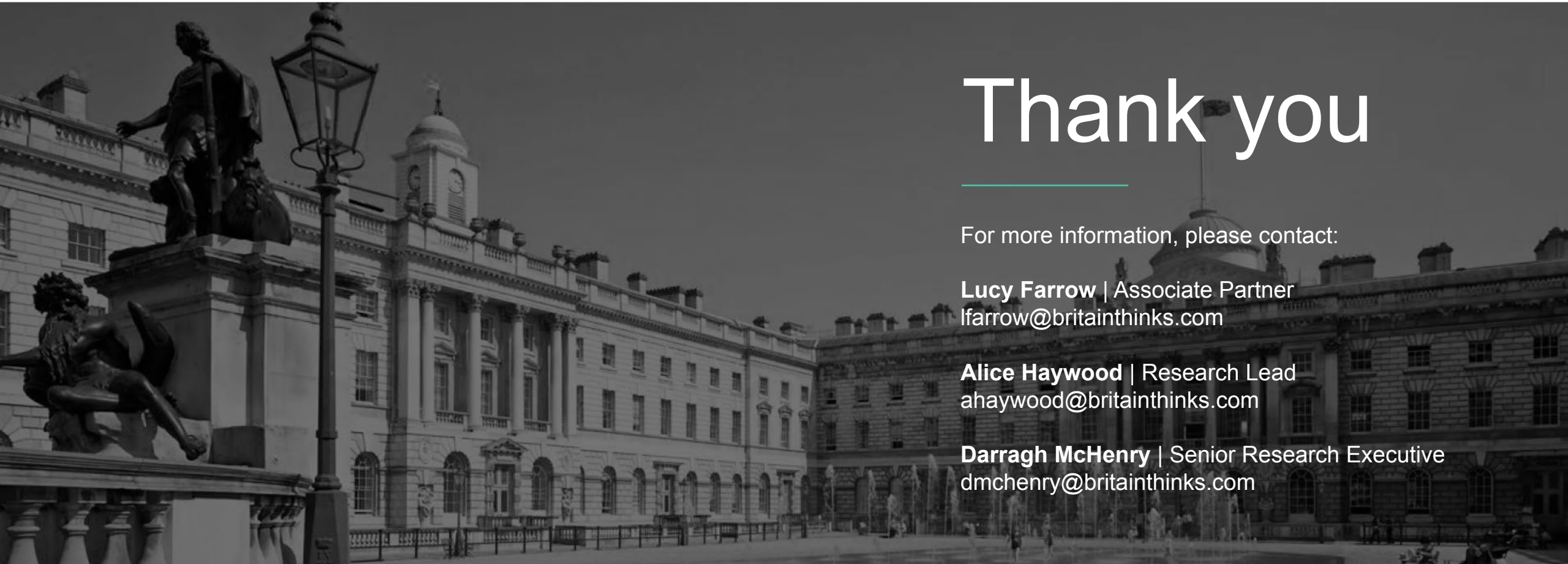
They assume that:

- AI will not be used unless it is performing a task at least as well as a human, if not better
- AI *should* follow certain regulation that is already in place, for example GDPR and anti-discrimination law, though there is some doubt about whether this will happen given the amount of personal data that current AI is already perceived to use
- AI systems are simple enough to explain in layman's terms, particularly explanations around the criteria used by these systems to make decisions

While the public don't have a clear sense of what kind of governance is needed for more advanced uses of AI, it is important to them that development of AI is well governed. For participants, this means respecting privacy, transparency, fairness and accountability for all users.

Participants recognised the limits of their knowledge, and wanted to see government and others acting on their behalf based on their fuller understanding of the issues.

This implies a need for proactive governance of AI which is in line with the principles the public support even where they are unsure about the application.



Thank you

For more information, please contact:

Lucy Farrow | Associate Partner
lfarrow@britainthinks.com

Alice Haywood | Research Lead
ahaywood@britainthinks.com

Darragh McHenry | Senior Research Executive
dmchenry@britainthinks.com