



Review into Artificial Intelligence

A research report from Deltapoll

July 2019

Background and methodology

Artificial Intelligence (AI) has the potential to rapidly transform public sector decision-making structures, with advanced computing and data science algorithms capable of identifying patterns within a myriad of data points to help predict and secure more informed, efficient and superior outcomes than those solely determined by human capabilities.

Like any new technology that has real implications for people directly affected by the decisions it arrives at, implementation of AI must be subject to the same standards, values and public expectations that underpin the society in which we live. That AI solutions can invoke complex moral and practical considerations reinforces the importance of this: accuracy, fairness, transparency, free of unintended bias and explanation are but some key principles that must be complied with to secure public trust in a future heavily featuring AI solutions.

Recognising the potentially far reaching implications of data-driven decision-making permeating public institutions, The Committee on Standards in Public Life announced a review into whether existing structures and regulations can ensure high standards of conduct. As part of the review, Deltapoll was commissioned to undertake two focus groups that investigated public attitudes toward artificial intelligence and public standards, with emphasis on how the Seven Principles of Public Life can be applied to AI and what ethical safeguards the public would want from it.

The core themes that the Committee wished to explore included:

* + How the provision of an explanation – or lack of one- into an AI derived decision would determine the extent to which it was accepted;
  + Who should be accountable for an AI decision, particularly in cases where detriment applies;
  + How unintended bias might impact on trust in AI decision-making;

Deltapoll conducted two focus groups in Central London on 13th June 2019.

One group was conducted among a ‘representative’ set of participants, reflecting the gender, age and ethnic profile of the local community. The second group was comprised of front-line public sector workers, including teachers, social workers, doctors, and members of the police force.

Qualitative research is an interactive process between researcher and participants: its strength is in allowing participants’ attitudes and opinions to be explored in detail, providing an insight into the key reasons underlying their views. However, discussion results are based only on a small cross-section of the public and public sector representatives, particularly so when budgetary considerations limited the number of these focus group sessions to just two. As a result, both Deltapoll and CSPL recognise the limited nature of the qualitative investigation, outcomes from which (reported here) are indicative rather than a full evidence base. Further, the design features employed imply that outcomes cannot be taken to be necessarily representative of the wider population.

However, Deltapoll also added a small set of bolt-on questions to our quantitative omnibus survey, comprising a nationally representative sample of 2,016 members of the Great Britain general public aged 18+. Interviews were conducted online on 14-17th June 2019, and results from these questions are weaved into this report.

Deltapoll is a member of the Market Research Society and the British Polling Council and complies with the rules of both organisations.

Research findings

|  |  |
| --- | --- |
| What is AI: potential, and fear  That professionals working in public-facing front line services are more aware of AI than their counterparts in the wider general public, and indeed are more thoughtful about the possible implications of its widespread implementation is probably no surprise. Nor should be the apparent ability of the latter group to catch up and come to quite strongly held and perceptive views about what AI might mean for them in the future.  AI is a clear and present - indeed exciting - proposition for public servants, being already recognised as a sophisticated resource for delivering expert decisions that extend decision-making possibilities beyond those that the human mind is capable of. There is a strong recognition that in a policymaking environment in which finite resources are - and will continue to be – stretched, a resource that is at least an aid to decision-making will be of considerable help. This is particularly thought to be the case in healthcare resource management, with prevention and diagnoses at the forefront of expectations. That is not to say that countervailing concerns around diagnoses exist among clinical staff, but that the opportunities are not only recognised but thought to have material, beneficial consequences.  Levels of understanding about AI were slightly lower in the general public group, but that is not to say they were either non-existent or for the most part inaccurate. There are misconceptions but perhaps not those that might have been expected. When presented with images of Arnold Schwarzenegger as The Terminator and I Robot sentient beings (Will Smith movie) for example, few people visualised AI as actually being or indeed intending to be like that. In short, nobody fell for the idea that AI is a Hollywood movie construct that inevitably ends with a battle of humans against the machines.  Rather, misconceptions focussed on the line between purposes that AI currently serves, and purposes that it doesn’t. The use of robotics during operative medical procedures was one incorrect assumption about the use of AI, while another was shopping in a physical context - where AI was presumed to be able to cut queuing times at checkouts. The fact that self-checkouts exists without an AI function didn’t occur, but once this view was articulated it strangely caught on among other participants who tended to agree with the point. However, this discussion pointed participants into more relevant applications of AI in the shopping context, with emphasis on online shopping and tailored advertising. AI recognition of previous shopping patterns to predict relevant products that people might want to buy now is largely seen as an irritation, either because the product previously looked at has now been bought or is unwanted. In acknowledging this common experience, the group intuitively understood that AI remains in its formative stages.  At this (still early) point in the public group discussion, a tangible crossover from limited experiences to more sophisticated appreciation of potential applications emerged. One line was the legal profession utilising AI to assist with due diligence on commercial contracts, or more general ways in which raw data could be quickly adapted to legal advantage. Another was the role of digital assistants in the home, such as Alexa, or on a phone (Siri).  While the advantages to the legal profession of an AI-based number crunching solution posed no ethical difficulties for the participant who mentioned it, recognition that digital assistants were ‘scaping’ personal data in apparent attempts to improve performance was both of itself a concern, and led directly to more general concerns about potential invidious uses of AI.  First of all, the very idea that computer systems can assimilate masses more data than a human and learn from it profoundly affected some people, who were quite simply scared by the prospect. Perhaps the more simplistic component of this fear was the loss of ‘humanity’ - engagement or other aspects of the human experience that people take for granted via normal interactions, but which will be lost in an AI future. (As an aside, the popularity of online shopping and the decline of high street retailing occurred to nobody as a current manifestation of this fear; seemingly, loss of human interaction is only to be feared if there is detriment rather than personal advantage).  The second component of fear was the potential for mendacious outcomes, deliberately constructed. While some people are ‘known’ to be bad and do bad things, concern does exist that computers can ‘learn’ to also be bad whether via the injection of scurrilous code by a programmer or simply by the machine becoming so clever through its own learning it can become an agent of harm. While this might be classic ‘Terminator’ territory that is informed by worst-case scenario fiction, the fact that people view AI as capable of becoming so dominant that it can take unilateral and incontestable decisions is evidently not.  Thirdly, and undoubtedly of greatest concern was the possibility of malfunction. Again, a malfunction was set against human ‘malfunction’ or error, which everyone recognises happens all the time, but there was a sense of extreme unease about the ability of an AI-enabled super-computer going wrong for whatever reason. This was largely because of the already established mindset that the power of human error was miniscule when set against the power of an AI failure. The wide raging implications of such an eventuality were not lost on people: what if AI includes false data that sends it off in wrong directions? What if banking systems are crashed as a result of AI malfunction? This particular point was a serious realisation – the idea that a banking malfunction could expose individuals or indeed the entire banking setup to unlimited chaos had the effect of seriously concentrating minds’ on the catastrophic consequences of exponential and uncontrolled AI development: ‘What ifs’ are a major point of concern to people when the system becomes cleverer than people.  However, discussion about negatives does not take place in a vacuum, and participants were equally assertive about the positives that could quickly emerge. The front-line workers group were especially cognisant of the true AI process, where machines become very good at understanding patterns in large amounts of big data and then take it one step further into recommendation setting. With healthcare being front of mind, the NHS administrator and the doctor were particular focussed on how it could assist with preventative and diagnostic decisions, with the real benefit of saving lives as a result. This was both in terms of improving medical science and in terms of increased productivity. With public services short of thousands of staff member ‘plan B’ was thought to be making best use of AI to increase productivity in a future continually featuring staff shortages.  The general public group, as might be expected, saw more simplistic but practical implementations for the general good. Google, for example, is already seen to have improved people’s lives in terms of information provision and will probably help in even more advanced ways in years to come. One such suggestion was that product prices will be tracked for individuals and a recommendation to buy will only be made when the algorithm learns what the perfect price point is in the era of dynamic pricing.  Other ideas revolved around the possibilities of a more empathetic humanity, with AI helping to solve homelessness, poverty or even prevent loneliness as future iterations of Alexa learn to be compassionate companions to the elderly, or even assist with household chores that it observes need doing (although this may return us to futuristic impressions of sentient robots).  Perhaps encapsulating all of this conversational entryism into the theoretical new world of AI was a headteacher’s juxtaposition of hope and fear. There appeared to be a heartfelt desire for it to successfully satisfy and close serious problems that exist in his world today. And yet, both understanding of what it could do for him, and how it would not scare the life out of him were conditions that had not been met.  Case study 1 accountability: Seven-year-old Alex and his AI formulated care plan  In the first of a series of (three) iterative case studies which highlighted both the potential and the ethical considerations associated with AI interventions, the provision of a care plan for seven-year-old Alex kick-started detailed discussion.  In this scenario, Alex showed signs of having special needs and had a care plan devised for him by his local council. When his parents met the council, they were surprised to find that the plan had been created via AI, and called for monthly counselling for anxiety. The Council could not explain why the software had reached this decision and didn’t know how the complex AI system worked.  Initial reactions to AI being used to ascribe care planning for a young person had to contend with natural human reluctance to accept that an individual human’s personality, behaviour and psychology could be deconstructed into numbers and code, and then ‘repaired’ by its decisions. There was a strongly held view, particularly among parents and senior teachers who had real world experiences of such things, that every individual is unique and only ‘contact’ experiences can inform decisions of such sensitivity and individuality. This was a view strongly upheld by a social worker, who believed that ten children who look similar on paper might have very different needs in reality.  This apparent rejection of AI judgements might be thought to be a classic reaction to a common misconception. To illustrate by analogy, in opinion polling people who decry polling because they have never themselves been polled often refuse to believe that a 1,000 or so people that they have never been part of can truly be representative of everyone. It can, and in effect, there is a failure to understand probabilities of selection and basic sampling theory. Similarly, in Alex’s case, there is a failure to grasp commonality within, and extrapolation from data. The idea that computer algorithms could as a matter of course ‘beat’ human experts is one that will likely be slow to catch on.  In addition, there may have been some emotional reactions to this idea on a personal level, given that those with some personal or occupational experience possibly felt threatened by technology taking over their responsibility and power of judgement. For others, there was a more balanced perspective though. Another teacher recognised that AI could be at least as accurate as humans in determining care plans, and might result in tangible time-saving outcomes.  There are two further caveats to the potential success – or perhaps acceptance – of AI care plans. Firstly, any care plan, irrespective of how it might be created, is thought likely to be questioned or rejected by parents if it fails to match their expectations. In such circumstances a second opinion might be inevitable. Secondly, and putting aside the objections articulated thus far, it was felt important for the ‘believability’ of AI as a potentially superior method of creating care plans that it must truly learn from its own conclusions - if it doesn’t then it is no more than aggregated data. Tangentially, this was thought to be important as a way of offsetting inevitable human failure or mistakes – a subject that was often alluded to but rarely taken to its natural conclusion that humans might not be as good as people think in making the right decisions. As one participant recognised, the public have been conditioned and socialised to put trust into a human, but not into code and computers. Trust in AI will not be forthcoming until it proves that works, even though trust is placed in humans despite inevitable mistakes.  The limitations identified immediately above were but preliminary and limited though. The main problem with the AI derived care plan was the failure to provide an explanation for the decision. There was no escape from this hugely problematic issue, which should be considered to be a potential red line that most participants were unwilling to cross. There were two basic reasons for this: firstly, that people need to trust the decisions that they receive, and secondly, an explanation signifies accountability i.e. that the buck stops with someone or somewhere.  As far as the former is concerned, much lies in the fact that the subject is medical diagnosis. Understandably, people with a condition want to understand as much about their (or Alex’s care plan) as they can in order to accept it, and the basic start point for understanding a plan is having the recommended course of action explained to them. In the absence of explanation, the danger becomes - as one participant indirectly articulated - that AI outcomes become joke-worthy: computer says no.  More seriously, from both the perspective of the recipient and of the provider, there is no obvious responsibility for the decision. One participant remarked how this would contrast with elected officials, all of whom are subject to the next evaluation of voters at the ballot box. In the AI scenario though, the prevailing view was that there would be nobody who could take responsibility if things went wrong.  The Alex case develops, with a different care plan produced by a human team, which isn’t explained in a way that Alex’s parents understand, but they accept it over the AI plan. At a later date, a software upgrade allows the system to output an easy to understand explanation, which is shown to Alex’s parents.  At this point in the life cycle of AI, particularly in relation to medical plans, the default position is to trust human analysis rather than AI analysis. Perhaps a genuine respect for authority or for the NHS underpins this, but there will be scepticism about the ability of AI to produce effective plans (in comparison to human created plans) until there is robust evidence to prove that the AI plans are better. Members of both focus groups appeared to moderate their objections if or when this were the case, although it remained the case that an explanation has to be provided for any care plan to be accepted.  It should also be remembered that even in scenarios where all parties understand the rationale for a care plan, a feeling that the current generation will be guinea pigs for its implementation still exists. It seems unlikely that there will be public acceptance of AI care plans until the evidence in favour of them is overwhelming.  So, what do the public think of this scenario? In conveying the basics of Alex’s case only one in three people said they would be ‘comfortable’ with AI generating a plan, with half the public (51%) uncomfortable with it. This rather fits the pattern of the focus groups, with more general unease with the prospect. However, there is a tangible generational divide on this, with young people much more inclined to accept AI than their older counterparts. For example, 54% of 18-24s are comfortable with an AI care plan, compared to only 27% of over 65s. There is undoubtedly more awareness and perhaps understanding of the potential of AI among this first generation who will grow up with it, which is a similar pattern to those we have seen on other new technological developments such as use of the Internet.    Parallel accountability case study: AI in the judicial system  In this alternative case study, AI is introduced into the judicial system, with the software assisting with judging if prisoners should be released for parole. The AI software analysed the criminal records of past prisoners, and then examined which prisoners, when released, went on to reoffend. It predicts which current prisoners would be likely to reoffend based on a 0-10 scale with 1 meaning they are certain not to reoffend. A judge uses the system and wants to grant parole to a prisoner with a score lower than 3. However, the software cannot explain its reoffending scores.  Objections to AI in this context are numerous and definitive. Primary amongst them is the conclusive nature of the decision, which is a black or white: freedom or no freedom. This contrasts strongly with the case of Alex, where a choice could be made by parents involved. Here, the prisoner has no say, despite their own liberty being at stake.  The implications of this decision were also thought to be too profound – potentially for the public as well as the prisoner. It is one thing for someone’s freedom to be evaluated by a computer, but another if incorrect decisions have an impact on public safety. Some prisoners are presumed to be dangerous, with an incorrect decision possibly resulting in direct harm to others. As such, it was felt that a decision of this nature and importance cannot be taken out of the hands of the experts.  But there were also principled objections about AI and how it might affect justice itself. There is presumed to be some legal certainty in the way in which people are judged, with for example, juries being required to evaluate whether a case has been proven or not. In our parole scenario, the lack of explanation for an AI decision implies that the reliance on solid evidence is subverted. In the extreme, failing to provide a reason to a prisoner for why they are not to be released could be rationalised as undermining natural justice. This also links into the lack of explanation argument that characterised the Alex scenario – without an explanation it is difficult to know who to blame if reoffending occurs.  Interestingly, in this scenario there were also two-way suggestions about how prisoners might game the system. Members of the law enforcement community present easily identified numerous occasions where one conversation had the power to change a life, while in the opposite context role model prisoners were released only to reoffend the next day. In either circumstance there was almost bewilderment that AI would be able to identify prisoners of either type, with the system compromised and untrusted as a result.  The general public strongly vindicate the thoughts of the focus group. Two in three people (67%) said they were ‘uncomfortable’ with the idea of AI making parole decisions, with women (69%) and the elderly chief amongst them. Out of all the potential contexts for using AI asked of the quantitative sample, the greatest reluctance was in this case scenario. It would seem that the public are just not ready for AI use in this way.  Case study 2: cataracts and accountability  A patient has an eye infection and visits the doctor who sends the patient for a scan. The doctor looks at the scans and also gives them to the IT team, which uses AI to process them.  The AI results suggest there is early stage cataracts. The doctor does not know why the AI says this as no signs of cataracts are visible. The AI has a known 90% accuracy rate, but human doctors only have 80% accuracy.  The doctor recommends that the AI result should be followed, and the patient should have eye surgery.  In this scenario, a much more positive response to AI was forthcoming. Part of the reasoning for this is that AI is looking at scan results, implying that its analysis is based on real and actual evidence. In effect, participants disentangled AI solutions based on cold, hard evidence (which they understood and could support) from those subjectively arrived at (where personal emotion and individuality also complicate decision-making). As such, using AI in this way was classified as an entirely more positive proposition.  Further, it was evident that greater trust is put in machine-read scans than in human diagnosis. This is perhaps due to scanning being an everyday home or office routine for many people, which they can understand from a practical perspective and relate to – they see the outcome of scan when it emerges from the photocopier machine and trust that it looks like the original. Even in a medical context, people can easily relate to what the AI is looking for and doing in the scanning process. They recognise that AI investigation of the scan might easily be superior to naked eye analysis.  Once again, the focus group reaction was mirrored in the survey results. In the case of medical scans and cancer diagnosis, over half (53%) were comfortable with the use of AI. Only 36% were uncomfortable with it. Further, there are few demographic variations in response, with men (57%) only slightly more comfortable with it than women (49%) and hardly any variation at all by age. It would appear that where tangible positive outcomes can be derived from the application of AI, the public are tempted by it.    In contrast to the previous scenarios, an explanation is auto-built into this diagnosis. Even though the doctor cannot see the evidence that the AI can, people are evidently much more comfortable when AI is objectively sifting through physical data patterns which can be articulated easily to the patient. If it finds cataracts from a physical scan, then cataracts it is assumed to be.  However, from the doctor’s perspective there are issues to contend with. Firstly, cataracts were classified as a relatively easy condition to judge, even if human diagnosis of it is not always correct. In a situation where AI saw cataracts, but the doctor did not, the medical instinct would be to recommend a period of observation rather than instant referral for a procedure. This might be because cataracts are not a life-threatening condition, making the risk assessment less urgent. With a more sinister potential condition though, the instinct might well be very different. It follows then that medical recommendations based on AI might simply become about risk management (and costs) rather than diligent application of AI conclusions.  Secondly, there is also a strong likelihood that clinicians will be – or will feel – undermined by AI diagnosis that goes against their own judgement. This is presumed to be a stronger possibility in relatively simple diagnostic situations like cataracts, rather than in more concerning situations, where multi-disciplinary teams would be more likely to be involved.  There are also some concerns from the public perspective. The so-called ‘false negative’ scenario is one – where perhaps AI does not pick up cancer, which is there, but the doctor accepts the negative AI conclusion to the obvious detriment of the patient.  The other is where an eye is operated on but found not to be suffering from any condition. This obviously links to the central point about AI medical diagnosis: who is to blame? In the first instance, there would appear to be more sympathy if multiple medics involved in the diagnosis, but little to none if this were a computer decision.  The surgery is performed but it turns out the AI was wrong; no cataracts were present.  The patient wants to hold someone accountable.  Who is to blame?  In actual fact though, considerably more sympathy was on view after consideration of the outcome. This is not only the case as far as the (very reasonable) victim/patient is concerned, but also in terms of patient complicity; there is an expectation that the patient must take personal responsibility, particularly if the situation was fully explained to them and/or they signed a consent form.  So, in advance of discussing who should be held accountable for the AI’s mistake, it is important to recognise that the risk was known (90% accuracy rate), explained in advance of surgery and a form probably signed. In searching for blame if there is blame to be handed out, then it is first to be found at the patient’s door.  But blame allocation after that is a quandary with no outcome. In a situation where the doctor, the software, the software developer, the NHS procurement team or indeed the hospital trust could all be culpable, no culpability could be reasonably found. The doctor is certainly not on the hook, and considerable sympathy was on display for the situation in which she found herself. There was some debate that the software should contain more data points to facilitate accuracy, and that the developer might have been negligent in some way (remembering that computers should recognise scan results) because there ‘should be’ 99% accuracy in cases like this. In reality, there was insufficient agreement to determine who was to blame. Indeed, it is fair to conclude that accountability is a hazy concept with AI – one that might well be too complex for the public to arrive at a considered and agreeable position on.  Although positivity in the use of AI in medical diagnosis was mitigated through discussion somewhat, it remained, on balance, a desirable long-term proposition. The feeling was that AI should be an aid to decision-making rather than the decisive influence on it, and that there should be a watching brief given that we are still in the early stages of its development. There was also partial support from the doctor present in the groups, who felt that its application should be focussed on more exploratory rather than clinical decisions. For example, rather than diagnosing cataracts it should be used to identify which antibiotic would be likely to have most success in dealing with a certain bug – something that a doctor could never quite know.  Case study 3: when AI is biased  The local police have developed ‘predictive policing’, which analyses personal data of every past drug dealer in the last 20 years to create a profile of a hypothetical drug dealer today.  The software then scans the profile against public databases to identify potential suspects. The police then use this data to obtain a warrant for a search. It has a remarkable success rate – convictions of drug dealers increase by 40%.  Given the presence of one police officer and one detective in the front-line focus group, the level of instant appreciation of potential downsides was tangible. However, immediate recognition of the main problem was also present in the public group: both sets of respondents immediately recognised that the system might well be discriminatory. The subject of race was the obvious primary consideration. It was thought that AI would simply replicate racial stereotypes with the impact being felt by the Black community, a disproportionate number of whom might suffer from unwarranted intrusion in their home in the search for drugs.  Not only was the use of historical data likely to be discriminatory, but it was thought to be largely unhelpful. Drug dealers today are not thought to reflect profiles of drug dealers of the past, with county lines cited as one example of how the drugs trade has changed significantly over the last few years. The input of old data on new variations in crime via AI was considered to be potentially misleading.  The question was also raised as to whether this was another example of the legal system being undermined or subverted without proper consent. If there must be probable cause in order to get a search warrant, the idea that AI profiling justifies it was seen to be tenuous at best.  A couple of developments were added to the case study, with the admission that the AI is biased against Black people, and then the discovery that it is only biased because of the data inputted merely reflected the racism of past police officers.  None of this came as any surprise. Indeed there was no need to add these developments because they had already been priced into the discussion. Rejection of AI in this scenario was uniform and universal. It was seen as dangerous, no less.  Survey respondents were asked about gun and knife crime rather than drug dealing, and maybe for that reason their view on using AI to profile for crime was more sympathetic. Although well short of a majority, four in ten (40%) would feel comfortable in AI being used in this way (although only 12% were ‘very’ comfortable with it while 28% were only ‘quite’ comfortable). More were against (47%) with almost as many (21%) very uncomfortable as quite uncomfortable (26%). Discomfort was higher among women and older generations, fitting the same sort of patterns that the poll revealed on other uses of AI.    In a parallel scenario on bias, AI was used by a local council to screen out unsuitable candidates seeking work based on their CV. It analyses previous CVs and learned about the characteristics of highly qualified candidates. However, it turns out that the system discriminates against the CVs of women.  In this case there was less immediate rejection of AI use, although a tangible gender divide on the case study was observable. Female group members were clearly uneasy with one male member who thought that use was justifiable, so long as gender was taken out of the analytical process; one woman explicitly rejected any form of profiling based on demographic characteristics.  The gender divide could also be explained by moral relativism associated with outcomes. The idea that CVs might be rejected was seen as much less serious than having a ‘door kicked in’ in the criminality scenario, but understandably the principle of bias (against women in this case) was a defining point for women in this discussion. This is borne out in survey responses. As the chart above shows, a majority (53%) would be comfortable with CVs being screened, but (not shown) is the variation between the level of comfort for men (59%) and the lower level of comfort for women (50%) despite gender not being mentioned in the question. We might speculate that female survey respondents would be much more uncomfortable if gender had been mentioned as a determinant of disqualification.  In other survey findings, comfort was on display for the use of AI in checking for fraud in immigration checks (64%). Indeed, this was by far the highest level of comfort in any of the scenarios presented in the survey, with medical scans and CV screening in joint second place (53% apiece). It might be the case that support for AI use reflects two things: one, that there has to be tangible and obvious benefits for society (cancer screening improved, for example), or that AI outcomes reflects some previously held form of belief or prejudice (support for immigration checks being seen as underpinning a populist view in support of falling immigration levels).  So, what’s the problem: Explanation, accountability or bias?  All three of the primary concerns with AI received considerable attention, were fully appreciated by focus group participants, and were robustly deconstructed in the pursuit of common ground. Perhaps the most obvious outcome was a firm rejection of AI if any kind of inherent demographic bias conditioned outcomes. This was reflected on both sides of the user vs ‘victim’ debate: the police would be as uncomfortable in justifying decisions based on demographic profiling as the public would be on the receiving end. That is not to say, though, that AI should NOT be used in the kind of ways the case studies focussed on. If AI can be made to be ***fair***, far fewer objections would be articulated. In the simplest of our contexts this implies that race is excluded from crime scanning and gender from CV application scanning, but in general any kind of systematic approach that incorporates stereotypes would receive short shrift from the public.  Linked to fairness is the requirement for accuracy. Few would see the point or value in the application of AI if it did not lead to better outcomes, although little to no thought was devoted to how ‘better’ might be measured. In effect, it came down to trust that AI would be able to out-perform human decisions, and that AI decisions would improve as it learned more and more about its subject.  Assuming that bias has no part in an AI application, then accountability becomes the focus. Whether it is because there was a failure to fully understand the true nature of AI itself, or because human involvement satisfies fairness considerations, there was a sense of agreement about the need for a **human** to be responsible for an AI outcome (even though we saw a surprising level of forgiveness for mistakes). This might well resolve many associated concerns: eliminating unfairness, reducing potential error, and increasing trust levels to the required standard. Human involvement and accountability may also be seen as a second opinion possibility, something that received frequent mentions in the cataract scenario.  In a strong sense, public acceptance of AI may simply come down to whether fairness is represented throughout the process, including most importantly, at the decision point. This is fully vindicated by the survey response, where the top two reasons for making people more comfortable with AI being a human always having a final say on the AI decision (69%) and there being a right of appeal to a human (66%).  Majority support is also there on most other factors that might be considered to instil further confidence in AI solutions. Having a 95% accuracy rate would help built trust (54% more confident as a result) and half (51% apiece) say that having an easy explanation and fully understanding how it works would increase comfort.    However, doubt does apply when it comes to the involvement of government regulators, with only 36% more confident in AI as a result of any contribution. This was also reflected in the focus groups, with a curt rejection in the idea of trusting the government with AI.  Nevertheless, the idea of AI itself is positively received, with obvious reservations and conditions attached. The focus groups did appear to accept the potential for life-improving contributions from various applications of AI, not only in terms of how it might assist medical and other processes but how it helps reduce impacts on scarce public resources.  *I won’t mind being a guineas pig, because at some point this system will work to a really good standard, and it will benefit people that come after that.*  **Male, public.**  *I think in some ways if we can get to a system where we can use data really well and not lose sight of the things we really need humans for then it’s all to the good because what’s the alternative really?*  **Civil servant** | *The more I’ve heard about AI, the more I think ‘actually, this is a step in the right direction’.* ***Police officer***  *I’m quite excited about the role that AI could play in delivering some of our public policy problems and answering some of the issues around predicting need, understanding demand and use of services.* ***Civil servant.***  *I am quite excited in favour of AI…in providing a more intelligent approach to issues such as preventing illnesses, public health and helping more front-line staff in terms of things like diagnosing diseases and morbidities.* ***NHS administrator.***  *Why would you physically need to replace a human with something that looks like a human, but isn’t human? I don’t see the benefit; you may as well just replace it with a machine. Or a computing system.* ***Male, public.***  *Being served in a shop by a sales assistant that is ridiculously slow. The fact that you could maybe do that with AI would be good….but on the negative side you’d lose the human interaction.* ***Female, public.***  *I do lots of online shopping and I think it needs to be a little bit smarter than it is now.* ***Male, public.***  *It frightens me, the thought of losing humans.* ***Female, public.***  *Anything that’s important, like life or death, I don’t think should be handed over to a machine, even if the machine is very clever.* ***Female, public.***  *There’s nothing to say that computers would not just take this information and just be encouraged to, I don’t know, there’s reason to presume that it’s bad things.* ***Male, public.***  *Does AI have different moods when it gets stressed? And then it just thinks ‘ oh bollocks, I’m just going to riot all this up’.* ***Male, public****.*  *Surely you could have a rebel AI? It must be possible.* ***Female, public***  *If our human brain is only a fraction of what the AI can hold, and the fact that AI could then malfunction, I’d be like, whoa. What could you do with that malfunction in comparison to what a human could do?* ***Female, public.***  *What if AI was handling your financial, banking, information, credit cards, spends. If that malfunctions the disaster it would cause ….surely the magnitude if that is so much greater than what one human could cause? I think that is my fear.* ***Female, public.***  *Being able to plug in lots of data about the population to make predictions about different illnesses affecting different people and then if it has a role in health prevention then theoretically it could be a cost saving and lifesaving thing for the NHS.* ***Doctor****.*  *I think the ability for a machine to or a robot to be able to interact with a person and give them more than their own voice or thoughts for that day, or some assistance with heavy bags is fantastic.* ***Female, public.***  *I’m very, very wary of it. It scares me. How would it help me and how could it possibly help kids? Someone tell me.* ***Headteacher.***  *Each child is individual, and you can’t base a care plan on the average of other people, it doesn’t work like that. I know that each child is completely different from another, and you can’t base one child’s needs on an average of another.* ***Female, public.***  *A computer….takes away the fact that I have got to know that child really well, so I can make a professional judgement”.* ***Headteacher****.*  *I firmly believe that the only people that can understand people are people.* ***Public sector male.***  *It’s like a tick box – this person’s got autism, he’s got anxiety, he’s got all this they make a recommendation based on that, there’s no way that’s going to be effective.* ***Public sector male.***  *I’m prepared to think it could produce something as accurate as human coding…it can free up for other things you need to do as a teacher.* ***Teacher***  *It would not matter if it were AI, GP or director of the entire surgery, if you’re not expecting what is delivered, then you’re not going to agree with it.* ***Male, public.***  *If you don’t trust in the fact that it’s learning then it’s just reproducing a collation of other conclusions. If it can genuinely learn then yes, fine.* ***Male, public.***  *I don’t think humans are always right actually, and they’re not always right in what works for a child so I can see AI triggering thoughts about possible treatment which then have a human filter.* ***Female, public sector.***  *It makes me think of that Little Britain sketch. The computer says no thing. Which is really frustrating.* ***Male, public.***  *I think my problem is more with the fact that they can’t explain why it is recommended.* ***Nurse.***  *You want the understanding of why the decision has been made before you trust it, I think.* ***Female, public****.*  *Where do you place the blame for these things? It’s like driverless cars, when it crashes, who do you sue?* ***Male, public****.*  *If AI is based on a certain % of success stories then OK, I would not see any reason not to try it. But I’m not walking out of a consultation from either one not understanding why this is happening.* ***Male, public.***  *It’s kind of still in its infancy and we just don’t know whether it will be successful until it did learn and if people come off certain plans and do improve, we might be more open minded.* ***Social worker.***  *I completely disagree with this.* ***Police***  *That’s saying’ freedom or no freedom’. This little Alex has, or his parents have an option. Like you can take this choice or that choice. You can use your judgement. There’s no judgement here, in this scenario.* ***Female, public***  *It’s a greater risk, I believe, because it’s the risk of everybody else and the public.* ***Female, public.***  *Our justice system is that you have legal certainty, yes? So things have to be explained to you so that you’re certain about why you’ve been treated in such a way by the system.* ***Male, public.***  *Are we going to say ‘oh the computer said we could release them’?* ***Police***  *What happens if the prisoner in question performs….throughout their sentence…that outsmarts the AI?* ***Female, public***  *I don’t know where it’s getting it’s data from or the information. I would not be happy to take that risk.* ***Police.***  *It’s looking at a scan which is an actual thing that’s been produced, and a machine can look at it in a different way than a human does.* ***Male, public.***  *I can see how it could scan a picture, read that picture, and analyse the picture. That makes sense to me. Being able to read someone’s behaviour, I think, is open to much more interpretation.* ***Female, public.***  *This kind of makes sense. It’s a scan. Let’s just say the computer software can look at that scan in incredible detail, down to the pixel, that the human eye just can’t see. So you kind of get how it works.* ***Teacher.***  *I would also in my head be thinking about what is the risk of not acting now?* ***Doctor***  *In this scenario, yes, I feel undermined….(because)…it’s more benign than say a really high-risk scenario.* ***Doctor***  *The bigger problem is when the AI is not picking up cases of cancer when it should be.* ***Police***  *I’d be a lot more forgiving of a multi-disciplinary team who’d made that decision than AI. I think I’d be really pissed off if the computers had got it wrong.* ***Civil servant.***  *I think I’d be quite forgiving actually, in this scenario, because I think we all make mistakes.* ***Civil servant***  *Your right. I have gone with that decision, haven’t I? The 90% decision. So I’ve got to take some kind of responsibility myself.* ***Teacher.***  *I’m not attributing any blame to the doctor in any way, shape or form.* ***Civil servant.***  *I think you understand though, nothing can be 100% correct. Neither human nor computer.* ***Social worker***  *If it’s more accurate than humans then I think there is a role, I’m not saying it’s the only role, but it definitely has a role.* ***Teacher****.*  *I fundamentally disagree with this one more than any of them. It doesn’t say what kind of personal data it has got. If you’re talking race….you’ve got disproportionate amounts of young black boys stopped.* ***Detective****.*  *I disagree with this. It’s discrimination basically.* ***Male, public.***  *Having past data is not going to be relevant today, at all. They are using 7 year olds on the county lines. It’s not going to be relevant.* ***Detective***  *This is contradicting the whole legal system. How can you use asset of data to obtain a search warrant? You have to have probable cause or suspicion or an actual act.* ***Male, public****.*  *Profiling anyone by biological, racial, physical characteristics I think is really dangerous.* ***Female, public.***  *I think the system works. It just needs tweaking to take gender out of it.* ***Male, public***  *But that’s the same as the other scenario. When you put comparables like race, sexuality, gender into it, I think that’s when it becomes biased.* ***Female, public.***  *It just needs to be fair. If it takes away the human decision it needs to be fair.* ***Male, public.***  *Going back to trust, this would work. Female, public.*  *It needs to be better than us.* ***Female, public.***  *Over time, this has got more chance of being fair than having 100% faith in a person being fair.* ***Male, public.***  *I just want someone to be held accountable for the final decision. If the computer makes the decision, who do you blame if something goes wrong? I just want somebody to oversee that because there can be room for error, just like we have room for error.* ***Doctor.***  *I think you should always have a right to appeal. You know, you can’t just be judged by AI. Otherwise it’s not fair.* ***Male, public.***  *I wouldn’t, no.* ***Male, public****.* |
|  |  |