MHRA AI and Regulation Engagement Report

Good regulation for AI in healthcare: what people with learning disabilities, carers and young people told us

LDE
Learning Disability England

carersUK

ayph

National Voices

Contents

Written by: Elizabeth Hoda, Louis Horsley

June 2026

1. Executive summary

National Voices was commissioned by the Medicines and Healthcare products Regulatory Agency (MHRA) to bring the perspectives of groups that may be underrepresented in standard deliberative processes into the National Commission's work on regulating AI in healthcare. This report summarises what young people, unpaid carers, and people with a learning disability told us about the conditions needed for safe, trustworthy and equitable use of AI in healthcare.

National Voices, in collaboration with Learning Disability England, Carers UK and the Association for Young People's Health, conducted three online workshops (2 hours each) with 26 participants recruited via our VCSE partners: 11 young people (16–25), 9 unpaid carers and 6 people with a learning disability. Discussions used three use cases of AI tools (AI scribe, AI booking/triage, diagnostic skin app) to explore how expectations change by risk and context.

Key findings

- **Human oversight and the human touch were non-negotiable.** Participants wanted AI to support, not replace, professional judgement, with clear routes to escalate concerns to a person.
- **Trust was conditional.** Confidence depended on demonstrable safety, clear safeguards and evidence of benefit. Participants prioritised practical assurances (clarity, consent, accountability and redress) over technical explanations.
- **Transparency and meaningful consent were foundational.** People expected clear disclosure when AI is used and plain-language information about what the tool does, what data it uses and what options people have.
- **Data use and commercial involvement drove anxiety.** Participants expressed concerns about third-party access, misuse and data breaches. Acceptance of data use for system improvement was highly conditional on robust safeguards.
- **Safety and ongoing scrutiny must keep pace with deployment.** Participants expected rigorous pre-market testing and active post-market surveillance, including the ability to fix or stop tools quickly if harms emerge.
- **Equity must be designed in.** Participants highlighted digital exclusion, accessibility needs and the risk of uneven roll-out widening inequalities.
- **Governance and accountability must be clear and navigable.** Participants favoured shared oversight across multiple bodies and straightforward routes for redress, with differing views on where responsibility should sit when things go wrong.
- **Different groups brought distinct priorities and concerns.** While there was broad alignment on core principles, young people often focussed on data use and commercial incentives; people with learning disabilities were the most risk averse group and strongly emphasised accessibility, autonomy and preserving human interaction; and unpaid carers often foregrounded

practical accountability, context-sensitive human oversight and fail-safe safeguards. Views within groups also differed on acceptable risk and pace of deployment, particularly amongst carers, which reflects the diversity and nuance of perspectives captured through this engagement.

If AI tools are deployed in the healthcare systems without visible safeguards, clear lines of accountability, and accessible routes to human support, it risks undermining trust and widening inequalities, even where tools could improve efficiency.

Key considerations

- Apply a **risk-tiered approach** to safeguarding with higher assurance thresholds for higher-risk use cases.
- Mandate **disclosure and meaningful consent** for AI use in direct care interactions.
- Require **public-facing, accessible communication** and support (not only technical documentation).
- Ensure **inclusive design and testing** and maintain **alternative pathways** that don't deploy AI tools to allow patients to 'opt-out'.
- Strengthen **incident reporting, redress routes and post-market surveillance**, including an 'emergency stop' mechanism.

2. Background context

Artificial Intelligence (AI) is increasingly becoming part of delivering health care services in England and internationally. Whilst regulation is evolving at pace, there are significant challenges in balancing speed and innovation with maintaining patient safety and public trust. To this end, the UK government has launched a National Commission into the Regulation of AI in Healthcare to advise the Medicines and Healthcare products Regulatory Agency (MHRA) and policymakers on a new framework for regulation. MHRA is the UK government body responsible for regulating medicines, medical devices and blood components for transfusion to ensure that these products are both effective and acceptably safe for public use. One priority for the Commission is to understand public views on AI to help design an effective and trusted regulatory approach.

As part of this, the Health Foundation conducted deliberative research to explore the UK's public views and the conditions necessary to foster public trust in the regulations. This involved three day-long and in-person workshops that convened members of the public to produce robust insights from a nationally representative sample. Whilst this will provide a strong evidence base to inform the Commission, this approach was not tailored to individuals experiencing time poverty, complex health and care needs or higher levels of AI scepticism.

Subsequently, National Voices was commissioned by the MHRA to undertake supplementary research. This focused on targeted engagement with communities that may be underrepresented in the primary deliberative dialogue: unpaid carers, young people and people with a learning disability. The aim of this project was to generate insights that complement the Health Foundation's findings and ensure that the perspectives of groups at risk of exclusion from standard deliberative processes were meaningfully included.

National Voices is the leading coalition of health and social care organisations in England, advocating for more equitable and person-centred health and care shaped by those who use and need it the most. With a membership of over 200 organisations spanning a wide range of health conditions and communities, National Voices connects policy and practice with the lived experiences of millions of people. Through this equity focus, engagement was delivered online to accommodate the needs of historically underrepresented groups as well as utilising established relationships with specialist voluntary sector organisations for an accessible, trauma-informed research design.

This report aims to capture the nuanced views and experiences of the three groups referenced above to inform the Commission's understanding of public perspectives on the principles underpinning AI regulation, and how trust can be built in a new regulatory approach.

3. Methodology

Phase 1: Scoping, discovery, and recruitment

To ensure alignment with the wider programme of work, we conducted initial scoping conversations with the MHRA, the AI Commission, and the Health Foundation. These discussions were used to refine the framing of the research and align question design with the Health Foundation's deliberative dialogue. This approach minimised duplication and ensured that findings from this research could be integrated effectively into the broader evidence base.

Participants were recruited in partnership with three specialist voluntary, community and social enterprise (VCSE) organisations: Learning Disability England, Carers UK and the Association for Young People's Health. In total, 26 participants were recruited across the three priority groups. Partner organisations were compensated through cofacilitation fees and recruitment payments in recognition of their expertise and role in enabling inclusive participation.

Recruitment focused on the following groups:

- Young people (11 participants): Individuals aged 16 to 25, with efforts to ensure diversity in gender and ethnicity, and representation across the four nations where possible.
- Unpaid carers (9 participants): This included A mix of younger (18–35) and older (55+) carers, representing a range of caring situations. Where feasible, recruitment aimed to include participants from across the four nations of the UK.
- People with a learning disability (6 participants): Individuals with mild to moderate learning disabilities, with a mix of genders and ethnic backgrounds. A smaller group size was intentionally used to support more meaningful engagement and allow sufficient time for each participant to contribute.

Phase 2: Co-production and materials development

Given the complexity of AI in healthcare and the existing trust deficit for some community members, a collaborative approach was adopted. VCSE partners were actively involved in co-designing the research materials and in co-facilitating the workshops.

This included a co-design session with all partner organisations to review draft discussion guides, identify potential trauma triggers and adapt materials to meet the specific needs of each group. The session also established co-facilitation approaches that centred community expertise, ensuring that the research process was accessible, inclusive, and responsive to participant needs. For example, the workshop with young people included digital tools such as online whiteboards and

polls, whilst the workshop for people with a learning disability reduced any heavy text and allocated ample time for group discussion.

Phase 3: Workshop delivery

All three workshops were delivered online and lasted two hours. Online delivery was chosen to reduce barriers to participation, including travel time and associated costs, and to enable participants to engage from familiar and comfortable environments. This approach was particularly beneficial for unpaid carers, who could remain available for caring responsibilities, and for participants who may feel less comfortable in formal or institutional settings.

Each workshop was co-facilitated by a National Voices researcher and a representative from a partner VCSE organisation. This model combined methodological expertise with community knowledge and established trust. Accessibility support was tailored to the needs of each group and included provisions such as a palantypist. Participants received a thank-you payment of £55 in recognition of their time and contribution.

The workshops used a structured discussion guide that allocated time for overview and context-setting, to ensure that opinions were informed, and were based around three use cases which were: (1) AI scribe that uses ambient voice technology to record consultation notes; (2) AI booking system that helps the patient book appointments by asking questions about symptoms and availability, and it decides how urgent the appointment is before booking; (3) diagnostic AI through third-party app approved by the NHS to take a photo of a mole or skin lesion for AI to assess and give a risk rating, plus advice on what to do next.

Structuring the discussion around these use cases helped explore the extent to which different AI tools impact perceptions and approaches to regulation. After each use case, there was a facilitated discussion on the key trade-offs to address the core research questions surrounding public perceptions on aims and outcomes of AI, trustworthiness in regulatory approaches and the potential safeguarding mechanisms needed for public confidence.

Phase 4: Analysis and reporting

Write-ups were produced for each workshop and analysis focused on identifying key themes, including areas of consensus and divergence, community-specific concerns and the trade-offs that participants found the most challenging. These trade-offs were categorised in the following manner:

1. Safety vs Innovation: Ensuring AI is safe without stifling beneficial innovation
2. Speed vs Scrutiny: Deploying AI quickly to address NHS pressures while maintaining rigorous post-market surveillance and safety. Once an AI tool is ‘approved’ and is in use, post-market surveillance is the ongoing checks and

reporting that spot problems early, learn from real-life use, and fix or stop the tool if needed.

3. Equity vs Efficiency: Making sure AI benefits everyone equally, not just those with digital access or in well-resourced parts of the country
4. Transparency vs Commercial Interests: Requiring explainability without revealing proprietary algorithms. Patients should be able to understand how the AI works without companies revealing trade secrets.

4. Key Findings

Initial perceptions and experiences of AI

Across the three workshops, participants came to the discussion with varied familiarity with AI, but a shared sense that AI in healthcare is hard to define, not always reliable, and not yet something they feel confident navigating without support. Young people tended to describe limited day-to-day use of AI, mostly for light-touch tasks such as study support, and were candid about scepticism, with one participant noting that "typically it's rubbish" (young people group). Young people's baseline trust in AI was also shaped by wider technology failures (including reference to the Horizon IT scandal) and concern that commercial incentives could override patient wellbeing.

Unpaid carers showed the widest range of starting points, from enthusiastic everyday use to strong reluctance, and several participants described using AI in practical ways outside healthcare. One participant captured this enthusiasm directly: "absolutely love it [AI]…and it's been absolutely amazing" (unpaid carers group). However, even those more positive about AI emphasised its limits and the continued importance of human reassurance and lived experience.

People with learning disabilities expressed the highest levels of anxiety and scepticism about AI being introduced without adequate explanation, often describing it as something being "thrown at" them. Participants repeatedly linked this to a need for basic, accessible education and genuine choice before AI is used in healthcare. One participant captured this gap in awareness plainly: "I didn't know any I never heard of artificial intelligence until last year which is poor really." (people with learning disabilities group) Another summed up a wider frustration with digital systems by asking: "why can't we just speak to a normal person?" (people with learning disabilities group)

Across all groups, these starting points informed to some degree what people needed to feel safe engaging with AI in healthcare: clear explanation in plain language, confidence that tools are used to improve care (not only to save money or drive efficiency), and the ability to speak to a person and opt out where needed.

Desired aims and outcomes of regulation

Across all three workshops, participants described regulation as the mechanism that should make AI in healthcare trustworthy in practice: ensuring people know when AI is being used, protecting privacy, setting appropriate safety thresholds based on the use case, and making accountability clear when things go wrong. Participants also emphasised that regulatory requirements need to be meaningful to patients, not only technical standards, so that people can understand what is happening and what choices and safeguards apply.

Young people framed desired regulatory aims around transparency, consent, and strong governance that prevents cost-saving or commercial incentives from

overriding patient interests. They repeatedly positioned AI as something that must remain under human control, with oversight that is shared across institutions rather than held by any single body. This was often expressed as AI supporting, not replacing, professional judgement: "AI should be a source but not the main source" (young people group). Alongside this, concerns about commercial access to data were prominent, including strong resistance to private companies handling identifiable health data: "I don't think any private companies should have, regardless of how they promise to use it and the rules around it...access to our health data." (young people group)

People with a learning disability tended to describe the aims of regulation in more immediate, practical terms: protecting people from harm in high-stakes settings, ensuring information is accessible, and guaranteeing that AI does not replace human care. Participants also emphasised autonomy and dignity as outcomes of effective regulation, including clear routes for support when technology is confusing or distressing. Their expectations were strongly tied to preserving the human relationship in care, with one participant stating plainly: "AI won't have the compassion" (people with learning disabilities group), and another reinforcing the wider view that human care should not be substituted: "no robot is as good as a human and no robot will ever be as good as a human" (people with learning disabilities group).

Unpaid carers consistently described regulation as needing to hard-wire human oversight, fail-safe escalation, and clear accountability into real-world pathways, particularly for complex situations where context, tone and lived experience matter. Carers also emphasised that AI should remain trackable and auditable, rather than introducing opaque decisions without clear routes for challenge or redress. One participant captured this expectation directly: "any AI needs some human tracking" (unpaid carers group). In terms of governance, carers favoured independent oversight involving a mix of bodies (not a single organisation), with one participant warning that a single body will "have their own agenda" and calling for broader accountability: "I don't think you can trust one body to be taking the full responsibility of it because they will instinctively have their own agenda at some point and that's what you want to avoid." (unpaid carers group)

Regulatory and governance approaches

Across the three workshops, participants emphasised that effective governance for AI in healthcare needs to be shared, independent, and accountable, rather than left to any single organisation. Participants also wanted governance arrangements to reflect the risk of the specific use case, and to avoid incentives that could lead to cutting corners. In the young people workshop, several participants supported a multi-stakeholder approach because they felt the NHS alone is "more likely to cut corners just to save money and resources" (young people group). This included independent experts and voluntary sector organisations.

People with a learning disability also tended to favour governance shared across bodies with different perspectives (including charities and patient groups) but were clear that accountability must be visible and navigable, rather than creating a system where people are bounced between organisations if ever a patient needed to engage with the regulatory structure. As one participant put it, whoever regulates must be clearly accountable and not "send us round in circles" (people with learning disabilities group).

Unpaid carers were similarly clear that governance should not be self-policing, and several participants focused on what accountability looks like when something goes wrong in day-to-day care. Some participants felt responsibility should sit with the clinician who checks and uses the AI output: "whose responsibility it was. It has to be with the doctor. It's their notes. And before you sign up on a letter, you read it to make sure it's correct" (unpaid carers group). Others framed responsibility more broadly as sitting with whoever is expected to check and assure safety, rather than attributing blame to technology: "I'd bring it down to whoever needs to check because at the end of the day I would not 100% put all my faith into AI ever. I would like to know that somebody, a person is there to check it through to make sure that dosage etc. that is correct" (unpaid carers group).

Trade-offs

Safety vs Innovation: Ensuring AI is safe without stifling beneficial innovation

Across all three workshops, participants supported innovation in principle but consistently treated safety as the first condition of acceptability. The acceptability of any given tool depended on the consequences of error, the ability to detect and correct errors, and who is put at risk if something goes wrong. This came through strongly when participants discussed the three use cases: an AI scribe to generate consultation notes, an AI booking system that asks symptom questions and decides urgency, and a diagnostic tool that assesses a photo of a mole or skin lesion and provides a risk rating before suggesting next steps. Participants were generally more open to lower-risk administrative uses (particularly transcription support) than uses that could directly shape diagnosis, triage or treatment.

In the young people workshop, this was strongly framed as avoiding "shortcuts" in an already stretched system, and concern that small error rates become unacceptable when the stakes are serious health outcomes. "There doesn't need to be shortcuts in the health system, the system is broken enough, and we don't need any more damage or risks to add." (young people group). This point stems from the concern that taking the risk on using AI tools to address capacity concerns could end up backfiring if an error occurred which resulted in time and resources being spent on investigations and compensation.

Young people's appetite for innovation also varied by use case. They were more relaxed about administrative support (for example, the AI scribe), where some felt "specialised transcription tools are not going to be noticeably less accurate than a

human taking notes" and could even be "more accurate than a person" in capturing appointment details (young people group). However, they were more cautious where tools influence clinical decisions, emphasising that acceptability depends on "how much power the AI's decision has over an overall decision" and that the risk increases if "they say it's this and no one's questioning it" (young people group).

The learning disabilities workshop drew the sharpest `red lines' around high-stakes contexts, especially medicines and diagnosis, with a clear expectation that the threshold for acceptable error should be much higher (and that people need evidence the tool works safely in practice). "We'll accept some minor mistakes, but when it comes to things like medications, it has to be 100%." (people with learning disabilities group) "If it's about meds, there should not be any mistakes." (people with learning disabilities group)

This caution did not always translate into rejecting AI outright. At least one participant described conditional openness combined with heightened vigilance: "I'm willing to give it a chance... there might be mistakes and teasing problems, but if it makes a serious mistake... I'm of the opinion I would go in, but I would go in wary." (people with learning disabilities group)

Among unpaid carers, safety was repeatedly linked to the limits of AI in capturing emotional context and nuance, especially in sensitive settings (for example mental health, end of life care and paediatrics), and the need for human judgement as a backstop. Participants tended to support a `fail-safe' approach where uncertainty or complexity triggers human review rather than an automated outcome. "AI has to be an assistant rather than the be all and end all." (unpaid carers group)

This connected to a wider emphasis that reassurance and human connection cannot be replaced by automation: "you can never, ever replace that one-on-one human interaction that's in front of you. We obviously want a voice of reassurance sometimes and AI will never have that lived experience." (unpaid carers group)

Speed vs Scrutiny: Deploying AI quickly to address NHS pressures while maintaining rigorous post-market surveillance and safety

Across all three workshops, participants accepted that the NHS is under pressure and were open to AI improving efficiency only if regulation and scrutiny keep pace. Participants consistently expected rigorous validation before deployment, plus active post-market surveillance once tools are live, including clear routes to fix problems quickly or stop use when harms emerge. They also wanted scrutiny to reflect the different risk profiles of the use cases: participants were more willing to tolerate limited errors in low-stakes administrative outputs if there is clear clinician checking, but expected far higher assurance for tools influencing urgency of access or diagnostic risk ratings.

People with learning disabilities most explicitly framed this as needing visible proof before trust is possible, and described trust as fragile, with serious mistakes likely to undermine confidence rapidly. "Need to see it to believe it." (people with learning disabilities group)

Young people emphasised that regulation is 'behind the curve' and expressed a strong expectation that governance should be established before widespread roll-out, rather than catching up after tools are embedded. Their scepticism was also shaped by wider technology failures beyond healthcare, including reference to the Horizon IT scandal, reinforcing the view that systems can fail at scale and that accountability must be clear before adoption accelerates. They also supported learning from errors (including storing and analysing mistakes for system improvement) as long as this does not widen access to identifiable data. This is especially true of giving access to private sector or third party organisations.

Unpaid carers were particularly practical about what scrutiny should mean after approval: clear accuracy benchmarks, routine clinician checking (especially for transcription-style uses), and an 'emergency stop' capability if real-world harms appear. This also linked to explicit debate about whether 100% accuracy is feasible in lower-risk contexts, with some carers willing to accept a lower benchmark where clinician checking is built in: "I think 100% [accuracy] isn't achievable. But that should be picked up because the doctor is going to check the notes and correct the mistakes." (unpaid carers group)

Unpaid carers also differed on the balance between speed and reassurance. Some supported faster deployment to realise benefits sooner, while others prioritised equitable access and accuracy even if that means a slower roll-out.

Equity vs Efficiency: Making sure AI benefits everyone equally, not just those with digital access or in well-resourced parts of the country.

Participants across all workshops treated equity and inclusion as non-negotiable, and often framed unequal performance or unequal access as a safety risk (misdiagnosis, misclassification, or inability to access care). Concerns included digital exclusion, communication needs (accents, English proficiency, speech impairments, sign language), and the risk of uneven roll-out creating a 'postcode lottery'. These concerns were also linked to how the use cases would be delivered in practice: if booking and triage becomes digital-first, or if diagnostic tools rely on data that under-represents some groups, then the groups already facing barriers could be disadvantaged further.

Young people were the most explicit about inclusive performance testing as a pre-condition for deployment, including testing across different accents and skin tones. Some participants preferred slower roll-out if that is required to ensure inclusive testing, while others debated whether phased roll-out could be acceptable if it delivers earlier benefit in some places. One participant argued for addressing

problems where possible rather than waiting for perfection everywhere: "surely you should just give it to them as soon as you can and not try and wait because it's better to solve some problems in some places, even if you can't solve other problems in other places, rather than trying to wait to solve all the problems at once, because you'll end up solving more problems if you just solve them whenever you can" (young people group). Others prioritised fairness and a co-ordinated national approach, reflecting concern about a postcode lottery: "if it's as groundbreaking and as helpful as we want it to be, then I feel like everyone should have it as soon as possible." (young people group) "[AI is] 100% required to be tested across every single race, gender, whatever else, every kind of skin tone to be usable or else you will kind of be misdiagnosed…[even] if it's going to take 15 years… It shouldn't be something we should do tomorrow… if it takes more time, then so be it." (young people group)

People with learning disabilities focused more on the practicalities of access - not everyone has a smartphone, app-heavy systems can be confusing or overwhelming, and digital-only routes risk excluding people who already face barriers. "There's too many apps now, too many technologies, but a lot of people don't have phones." (people with learning disabilities group)

Unpaid carers tended to frame equity through the risk of a two-tier system and the expectation that alternative pathways must remain, so that improved efficiency does not shift burden onto people who are digitally excluded or managing complex caring responsibilities. Many participants in the group with people with learning disabilities also wanted to maintain alternative pathways for patients who do not want or cannot use AI tools in their care.

Transparency vs Commercial interests: Requiring explainability without revealing proprietary algorithms and trade secrets

Transparency and informed consent were consistently treated as foundational to trust, particularly where AI is used in direct care interactions. Participants wanted to be told when AI is used (including in appointments), understand what it is doing in that context, and know what choices and escalation routes are available. They also prioritised accountability (`who is responsible if harm occurs') over technical explainability.

The young people workshop contained the strongest reaction to undisclosed AI use and private company involvement, framing it as a breach of trust and inconsistent with expectations created by GDPR and data protection norms. This was reinforced by an example one participant gave of discovering AI use only after the event (for example, realising an AI transcription tool had been used after receiving appointment notes), which they described as fundamentally at odds with informed consent: "it's a little bit insane that we have such strong regulations around GDPR and data protection and protecting your privacy… and then we just go and let random private companies use their services with your data and you've not

approved it and you've not been told about it, which I just think is absolutely insane and goes against everything that we all speak about all the time" (young people group).

Even where young people saw a case for using data to improve tools, they drew a strong distinction between anonymised and identifiable information: "it's important to have this data to train things, but so long as it's not linked with you in any way and it's all anonymous. I think that's totally fine to have but as soon as it you know you've got a name and whatever attached to it then that should absolutely be kept private from any external companies" (young people group).

People with learning disabilities emphasised plain-language explanation and simple, accessible consent mechanisms (including the ability to say no), linking transparency directly to autonomy and dignity. They were particularly clear that consent should be explicit and optional: "they should ask us if we want this, and if we say no, don't use it" (people with learning disabilities group).

Unpaid carers raised the most detailed concerns about the downstream consequences of health data being used beyond care, including marketing, insurance implications, credit and DVLA concerns. They also linked these concerns to the practical realities of caring relationships: where carers support someone else's access to care, participants wanted reassurance about safeguards against misuse, and clarity about who can access what information and why. This strengthened calls for strict boundaries on the 'data journey' (storage, access, retention) and on third-party involvement. "the last thing I want to do is have a health condition I want to keep that private in terms of consultation, but my worry is if that then goes to a third party, Am I going to get bombarded by advertising? Am I going to get penalised because my health condition now means my health insurance is going up… Is it going to affect me having a mortgage?... having a credit score." (unpaid carers group).

5. Considerations for MHRA to enable effective regulation

Whilst the considerations below are drawn from our workshops with young people, people with a learning disability and unpaid carers, we feel confident that implementation would go some way in increasing levels of trustworthiness in the use of AI in healthcare across all patient groups.

These considerations are grounded in the views and concerns expressed across all participant groups. Following analysis of the focus groups, the final considerations presented below have been informed by participants' perspectives and subsequently further refined by the National Voices team.

Theme: Trade-offs (Safety, Equity, Growth)

1. Adopt a risk-tiered approach to regulation with a universal baseline, plus stricter controls where risk for harms are higher.

On the whole participants did not treat AI in healthcare as one thing, but rather evaluated acceptability through the lens of what the specific application is – what the tool does, the risks and benefits, and who is accountable if anything should go wrong. More administrative tools (such as appointment booking) tended to be seen as less risky and therefore required less strict controls whilst maintaining baseline standards around safety, dignity and inclusion. A tiered model of regulation allows faster adoption for low-risk tools while requiring high evidence thresholds and stricter safeguards for higher-risk uses such as diagnostics.

2. Define minimum safety thresholds that are use-case specific.

Overall, participants judged acceptable safety performance by the consequences of error, the ability to detect and correct errors, and the person at risk if an error occurs. As outlined in the above, participants distinguished between tools that change clinical decisions and/or outcomes and tools that support administration. In general, a higher safety threshold was demanded for clinical and diagnostic contexts; whereas minor errors were more tolerable for lower stakes use cases. A regulatory approach that applies the same threshold to all tools could either over-burden lower-risk tools or under-protect people in high-risk contexts.

3. Make equity a condition for approval, not an afterthought.

Participants often linked bias and exclusion directly to harm because if an AI tool performs worse for some patient groups, then in the short-term it can lead to late or inaccurate diagnosis, inaccuracy in records or ineffective treatment. In the longer-term, exclusionary practices could exacerbate existing inequalities and widen the trust deficit. If a tool is approved and deployed widely before inclusive testing and accessibility adaptations are locked in, the system can normalise a two-tier service.

Developers should adhere to inclusive design principles and validate them in testing.

Theme: Trustworthiness

4. Mandate informed consent and disclosure where AI is used in direct care interactions.

For many participants transparency and choice were foundational conditions for trust in the use of AI tools, both of which require disclosure. Undisclosed use of AI risks undermining the relationship between patient and care provider, which can further damage trust. Undisclosed AI use also prevents patients from weighing risks that matter to them (privacy, third-party access, data retention, and the acceptability of AI involvement), and it disproportionately disadvantages people who need accessible information and clear options to exercise agency.

5. Require demonstrable evidence (pre-market) plus ongoing monitoring (post-market) that is visible to the public

Participants treated trustworthiness as conditional on evidence that AI is safe and fair before use, and on ongoing scrutiny once tools are live. Making both evidence and monitoring visible to the public was central to accountability and informed choice, helping patients understand what level of trust is warranted and enabling problems, including unequal impacts across communities, to be identified and acted on early.

Theme: Safeguards and Oversight

6. Requirement for human oversight of AI outputs in higher-risk decisions.

Workshop participants repeatedly treated human oversight as non-negotiable in higher-risk uses of AI (such as diagnosis), because they saw these tools as capable of causing serious and sometimes irreversible harm if they are wrong. Having a human-in-the-loop is especially important in cases where understanding nuance, context and vulnerability is key to providing good care.

7. Build 'fail-safe' escalation routes and an 'emergency stop' process into regulation.

Participants expect regulation to specify what happens when things go wrong in real-world use. Escalation routes protect access to human support when an AI tool fails or is confusing, and an 'emergency stop' creates a practical mechanism to pause deployment when safety, equity, or performance concerns emerge.

Theme: Liability, Accountability and Responsibility

8. Clarify accountability across the pathway (developer, deployer, clinician, system).

Participants described clear lines of accountability as a fundamental component of effective regulation. Being able to understand who is ultimately liable if something

were to go wrong would make some patients feel more comfortable about the use of AI in healthcare. This is especially true when health and care needs may be more complex due to the involvement of carers, or for patients living with multiple conditions who rely on coordinated support from various parts of the healthcare system simultaneously.

9. Require incident reporting, redress routes, and compensation mechanisms that are easy to use.

As well as clear liability, participants emphasised the importance of being able to access simple mechanisms to address issues related to their care that may arise from the use of AI in the pathway. Visible accountability, easy escalation and clear routes to redress would address some of the concerns amongst patients. Developers should also consider implementing a 'you-said-we-did' system to report back on what changes have been implemented.

Theme: Transparency and Explainability

10. Require transparency on where data is stored, who accesses it, retention periods, and whether third parties are involved.

Participants linked trust directly to understanding the 'data journey' and who benefits from access. Clear, standardised disclosure enables informed consent, helps people assess risks of misuse or onward sharing (including commercial access), and supports accountability if something goes wrong. This is especially important when a large trust deficit already exists, such as the presence of commercial organisations within the pathway.

11. Require that an explanation of how the AI tool works in any given context is accessible to the patient.

Across the workshops, participants treated understanding as a condition for trusting an AI tool. Understanding why and how an AI tool led to a certain decision or outcome are core to developing trust. Without accessible explanations, participants felt patients cannot give informed consent, challenge errors, or navigate escalation routes – making AI tools feel unsafe and exclusionary.

Theme: Access and Equity

12. Guarantee alternative pathways that don't require the patient to use digital-only AI-enabled pathways as a regulatory requirement.

Participants described how being able to access a person as both a safety net and a matter of dignity. Non-digital and human routes reduce the risk that AI-enabled pathways deepen digital exclusion, create barriers for people with communication needs, or leave people stuck when tools are confusing, inaccurate, or distressing. This is especially important in the early stages of adoption, but should remain an option indefinitely.

13. Ensure developers adhere to inclusive design principles (disability, language, neurodiversity, communication needs) and validate them in testing

Participants framed inclusivity as a pre-condition for safety and fairness, not a feature to add later. Setting standards and testing them reduces predictable harms (misunderstanding, misclassification, non-access) and helps prevent a two-tier service where some groups benefit while others are left behind.

6. Considerations for wider stakeholders to enable effective implementation

Some of the considerations that emerged from this research fall outside the remit of the MHRA but are still relevant when considering how to effectively implement AI tools in healthcare.

1. Require public-facing communication in plain language, designed for accessibility.

Participants treated understanding as a precondition for trust, consent, and safe use. 'Buy-in' will come from education and awareness raising. When people cannot understand where AI is being used, what it is doing, and what their options are, they are not able to make informed choices, challenge errors, or seek help. Providing jargon-free information and support to understand how AI is being used in any given context can lead to increased trustworthiness. This is also beneficial for people who don't speak English as their first language. If the use of an AI tool seems opaque, no amount of effective regulation or safeguards will make patients feel safe.

2. Guarantee alternative pathways that don't require the patient to use digital-only AI-enabled pathways as a regulatory requirement.

Participants described how being able to access a person as both a safety net and a matter of dignity. Non-digital and human routes reduce the risk that AI-enabled pathways deepen digital exclusion, create barriers for people with communication needs, or leave people stuck when tools are confusing, inaccurate, or distressing. This is especially important in the early stages of adoption, but should remain an option indefinitely.

3. Mandate staff training and clear operational guidance.

Participants consistently framed safety as something that depends on how AI is implemented and used day-to-day, not just how the tool is built or performs in pre-market testing. Participants also repeatedly emphasised that AI cannot capture tone, body language and complex context in the same way a clinician can, so staff need clear guidance on when to rely on AI outputs, when to escalate, and how to ensure the 'human touch' and reassurance remain available.

Appendices

Demographic		
Gender	Male	10
	Female	14
	Other	1
Ethnicity	White – British, Welsh, Scottish, Northern Irish	15
	White – Gypsy or Irish Traveller	1
	Asian or Asian British	5
	Black or Black British – African, Caribbean	4
Age	< 18	4
	18 - 24	8
	25 - 30	1
	31 - 35	2
	36 - 44	4
	45 - 54	2
	55 - 64	2
	< 65	2
Health condition and/or neurodiversity	Physical health condition	15
	Mental health condition	3
	Neurodiversity	6

*1 participant from the people with a learning disability group chose not to disclose their demographic information.

Acknowledgements

National Voices would like to thank our VCSE partners and communities who shared their valuable expertise and insights through the workshops conducted in the course of this project:

- Learning Disability England
- Carers UK
- Association for Young People's Health

National Voices

National Voices is the leading coalition of health and social care charities in England. We work together to strengthen the voice of patients, service users, carers, their families and the voluntary organisations that work for them. We have more than 200 members covering a diverse range of health conditions and communities, connecting us with the experiences of millions of people.

Reg. Charity 1057711
Company Ltd. 3236543

020 3176 0738
info@nationalvoices.org.uk
@nationalvoices.bsky.social

www.nationalvoices.org.uk

The Foundry,
17 Oval Way,