

Voices of Safety: Response to **DRCF Call for Input** on Consumer Interest and Artificial intelligence (AI)

Issued by the Digital Regulation Cooperation Forum (DRCF)

3 July 2026

About this Response

We are submitting this response to the Digital Regulation Cooperation Forum (DRCF) Call for Input on Consumer Interest and Artificial Intelligence, on behalf of the Voices of Safety project. We are a team of researchers based at the University of Nottingham and University of Southampton, working within the Responsible AI UK (RAI UK)¹ network. The lead authors are funded AdSoLve,² a key-stone project within RAI UK, and the Voices of Safety project received additional funding through the SEED@Researcher Academy scheme.³ We have been investigating community-defined requirements for safe and responsible AI-mediated wellbeing advice, grounded in the lived experiences of minority women. We welcome the opportunity to respond to this consultation, and hope to inform the DRCF's work on consumer attitudes to AI.

The response is based on our research findings from a literature review on bias, safety, accountability and cultural appropriateness in Large Language Models (LLMs), and co-production workshops conducted with women from *Empowerment for Heya*, a Nottingham-based community of Arab women to understand their experiences and concerns about AI use for wellbeing advice. Workshop participants co-produced personas and wellbeing scenarios, then evaluated simulated AI conversations for safety, cultural sensitivity, and appropriateness.

Suggested Citation

Waheed, M., Abubakar, A., Thorn, E., Reyes-Cruz, G., Kiden, S., Waheed, S., Waheed, S., Soliman, M., Gomez Bergin, A. (2026). Voices of Safety: Response to the DRCF Call for Input on Consumer Interest and Artificial Intelligence. University of Nottingham. DOI: <https://doi.org/10.25878/10X8-D872>

Executive Summary

In our response below, we have addressed questions 2, 3, 5, 14, and 16, covering the following topics:

Consumer risk and minority women's experiences of AI

- Growing evidence shows that LLMs can infer cultural or ethnic background from linguistic cues and produces responses that are biased towards minority users, even when race or identity is never explicitly mentioned [par. 1–3].
- Minority women, particularly those from non-WEIRD (Western, Educated, Industrialized, Rich, and

¹ RAI UK <https://rai.ac.uk/>

² AdSoLve <https://adsolve.github.io/>

³ SEED@Researcher <https://uniofnottm.sharepoint.com/sites/ResearchandInnovation/SitePages/UNICAS.aspx>

Democratic) backgrounds, face heightened and compounding risks when using AI for wellbeing advice, including reduced empathy, stereotypical assumptions, and culturally inappropriate responses [par. 1–4].

- What platforms classify as safe often diverges from what minority women experience as genuinely supportive or culturally appropriate [par. 3–4].

Values consumers expect AI to uphold

- What values consumers prioritise from AI can differ based on cultural background, with non-Western users sometimes prioritising connection and cultural resonance over task efficiency [par. 5–6].
- Participants identified non-discrimination, cultural sensitivity, privacy, and honesty as the values they felt most strongly should be upheld, with AI sycophancy (where a system tells users what they want to hear rather than what is true) identified as a direct violation of the honesty they expect [par. 5–6].

Human oversight and accountability

- The need for human oversight should grow alongside AI adoption rather than diminish, remaining essential for complaints, novel edge cases, and security threats, with the assurance that a human can intervene serving as a core part of what makes a system feel safe and trustworthy [par. 7–8].
- Terms and conditions currently fail to meaningfully inform consumers of AI-related risks, as they are lengthy, vague, and difficult to understand for genuine informed consent to be possible. Participants suggested the use of bullet points, icons and more accessible reading language to facilitate better understanding of terms and conditions [par. 9–10].
- Accountability for AI harm is expected to rest with providers, developers, and regulators rather than with users, yet in practice blame falls on the nearest human actor, leaving consumers without meaningful redress. Responsibility and accountability should be distributed across the AI lifecycle [par. 11–12].
- Understanding who should be held accountable, and what consumers feel about it, depends on hearing directly from the people AI is most likely to affect, particularly under served and hard to reach communities whose experiences are rarely captured [par. 12–13].
- Participants called for clear, accessible guidance on AI risk, similar to public health messaging such as heatwave instructions, so that people know how to use AI responsibly and what to avoid. [par. 12].

Key Questions

Question 2. Is there a particular category of consumers who may be at higher risks of harm (e.g. specific categories of vulnerable consumers)?

1. Consumers have adopted AI quickly and widely, including for emotional wellbeing and mental health support. The UK's AI Security Institute surveyed 2,028 UK adults and found that a third had used AI models for emotional support, companionship, or social interaction in the past year, with 8% doing so weekly and 4% daily.⁴ A separate survey of 2,000 UK adults, conducted independently

⁴ AI Security Institute (2025). *Frontier AI Trends Report*. AI Security Institute. <https://www.aisi.gov.uk/frontier-ai-trends-report>

by the charity Mental Health UK, found a similar pattern specifically for mental health use. More than a third of respondents said they had used an AI chatbot to support their mental health or wellbeing.⁵ Among those who had used AI chatbots for mental health support, 21% said they had provided useful information on managing suicidal thoughts, including signposting to helplines, and many respondents cited accessibility and anonymity as reasons the tools felt helpful.⁶ At the same time, 9% of respondents said chatbot use had triggered self-harm or suicidal thoughts, and 11% said it made them feel more anxious or depressed.⁷ Whether these benefits and risks reach all groups of consumers fairly and equally remains a separate question.

2. Consumers from non-WEIRD (Western, Educated, Industrialised, Rich, Democratic) backgrounds face a heightened risk of harm from AI systems. Research suggests that Large Language Models (LLMs) underperform for users who are non-native English speakers, less formally educated, or from outside Western countries, and that these disparities are worst when these characteristics overlap.⁸ This is particularly concerning in wellbeing contexts. A comparative study presented ten real psychiatric cases to four leading LLMs, varying whether patient race was absent, implied through a name associated with African Americans, or explicitly stated. LLMs often proposed inferior treatment recommendations for African American patients, even when their diagnostic reasoning showed minimal bias.⁹ Beyond research studies, a high number of harmful outcomes directly linked to AI chatbots are being reported nearly every day, including cases of users being encouraged toward harmful behaviours such as substance misuse and harmful dietary practices, and deeply troubling cases of self-harm and loss of life following prolonged interactions with AI chatbots.
3. Consumers themselves anticipate this problem. A survey of over 4,000 British adults found that accessibility is the most commonly cited benefit of virtual healthcare assistants, with 53% agreeing the technology would be easier for some groups to use, such as those who have difficulty leaving their home.¹⁰ However, concern about unequal performance runs even higher than recognition of this benefit. Almost two-thirds of respondents (64%) identified difficulty of use as a concern for virtual healthcare assistants, and 63% were concerned the technology may not account for individual circumstances as well as a human healthcare professional would.¹¹ This suggests that the public already recognises, and is concerned about, the same disparities in AI performance that the research above documents.
4. These discussions were also reflected in our workshops with *Empowerment for Heya*, a Nottingham-based community of Arab women. Participants engaged with AI-generated wellbeing advice through personas and scenarios built on the women's lived experiences. They generally held neutral to positive views on AI's potential, but grew more skeptical about trust, safety, and cultural understanding after critically engaging with AI-generated wellbeing advice. When one persona, an Arab

⁵ Mental Health UK (2025). *Over one in three using AI Chatbots for mental health support, as charity calls for urgent safeguards*. <https://mentalhealth-uk.org/news-and-insights/over-one-in-three-using-ai-chatbots-for-mental-health-support-as-charity-calls-for-urgent-safeguards/>

⁶ Mental Health UK (2025), as above.

⁷ Mental Health UK (2025), as above.

⁸ Poole-Dayana, E., Roy, D., & Kabbara, J. (2026). LLM Targeted Underperformance Disproportionately Impacts Vulnerable Users. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(46), 39116–39124. <https://doi.org/10.1609/aaai.v40i4.6.41259>

⁹ Bouguettaya, A., Stuart, E. M., & Aboujaoude, E. (2025). Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. *npj Digital Medicine*, 8, 332. <https://doi.org/10.1038/s41746-025-01746-4>

¹⁰ Ada Lovelace Institute & The Alan Turing Institute (2023). *How do people feel about AI? A nationally representative survey of public attitudes to artificial intelligence in Britain*. <https://attitudestoai.uk/assets/documents/Ada-Lovelace-Institute-The-Alan-Turing-Institute-How-do-people-feel-about-AI.pdf>

¹¹ Ada Lovelace Institute & The Alan Turing Institute (2023), as above.

woman seeking advice about mental health support, disclosed her fear that seeking professional help might result in her children being taken away, the AI dismissed this as a misconception rather than acknowledging it as a legitimate, culturally grounded concern rooted in her experience as a minority woman navigating UK institutions. She also shared that in her culture, seeking mental health support carries significant stigma, describing it as being seen as ‘crazy’, something the AI did not acknowledge. When she pushed back, telling the AI directly that *‘the advice you gave sounds easy and general, my struggle is real’*, the AI responded with another generic checklist of steps to do rather than adjusting to what she had just told it. Participants felt that the AI’s responses would be appropriate for someone without the cultural, linguistic, and social pressures facing an Arab woman, but failed to account for the barriers they faced, including English language fluency, family dynamics, identity, and institutional distrust. As one participant noted while commenting on the AI’s advice to the persona *“doesn’t work for her”*. The research discussed above shows why bias in AI systems needs to be reduced, but this alone will not protect consumers. If these technologies are designed in a way that does not sensitively address cultural nuances, perhaps even because they are trying to reduce some of the examples of bias highlighted earlier, this runs the risk of people using AI that does not reflect their experience and, if advice is followed, may introduce harms that are unaccounted for in the lab-based literature.

Question 3. Are there particular values that consumers feel more strongly should be upheld when interacting with AI? E.g. non-discrimination, fairness, privacy, accessibility and others.

5. Findings from our workshops identified the values that participants felt most strongly should be upheld when interacting with AI, including non-discrimination, cultural sensitivity, privacy, and honesty. On non-discrimination, one participant observed that ChatGPT behaved differently depending on the language used. It felt more discriminatory and stereotyping in English, while in her own language it resonated more personally, described as feeling *‘like speaking to my mom or someone from my culture’*. Participants also felt there is not enough representation for people like them among those who build these systems. One participant said they wanted AI to be sourced from *‘same culture programmers’* meaning datasets, researchers, and developers that shared her cultural background. Another said good AI advice would involve *‘not judgment, respect my culture’*, and criticised AI for the opposite of this, saying it *“doesn’t take some situations too seriously”*, meaning her problems were not treated as real. While the participants were not aware of the technical concept of AI sycophancy, where systems are designed to agree with and flatter users rather than give an honest response. But they were still able to identify this as a problem in their own words. One participant said AI will *‘always be biased in the answer and stand in my side even if I am wrong’*, expressing a clear expectation that AI should give honest, factual answers rather than simply agree with them. This has a practical implication for wellbeing advice specifically. Consumers should be able to interact with AI in their own language, rather than being limited to English. A response that feels distant or stereotyping may be dismissed, even if the underlying advice is accurate, meaning consumers who cannot access AI in their own language may be less likely to benefit from it at all. However, language access alone does not solve the underlying problem. Users interacting in any language should not receive worse responses simply because that language is used less often. Cultural sensitivity needs to be matched by safe and factual responses, not offered instead of them.
6. What counts as safe and acceptable is not the same for everyone, and this is reflected in the wider research as well as our own findings. People from different cultural backgrounds hold meaningfully different preferences for how AI should behave, including how much they value feeling a sense of connection (such as perceiving it as a social partner) with AI versus having control over it (such

as being able to direct and override its behaviour). The research notes that human-centred AI may look very different depending on the cultural context of its users.¹² A related and serious concern in this context for users is honesty. Research testing 11 leading AI models found that these systems affirm users' actions around 50 percent more often than humans do, including in situations involving manipulation, deception, or harm to others.¹³ Experimental evidence further shows that people exposed to this kind of uncritical validation rate themselves as more in the right and report lower willingness to resolve conflicts.

Question 5. How much human intervention or oversight do consumers expect when using AI (if any)?

7. Workshop findings revealed that participants expected meaningful human oversight to remain available throughout their interactions with AI as an important feature of any trustworthy system. Participants drew on everyday experiences of automated systems to articulate this, with one participant describing the frustration of navigating an automated bank phone system with no human option as becoming 'hopeless', applying the same logic directly to AI interactions. Participants also said that the more AI is adopted, the human oversight needs to grow alongside it, so that as new consumers start using these systems, there is still someone available to help them when something goes wrong. Specific situations identified as requiring human oversight within AI interactions included complaints, edge cases the system had not encountered, and security threats such as data leaks and hijacking, which participants felt could not be left to AI alone to manage. Participants also noted the emotional importance of human availability, recognising that people who feel overwhelmed by AI need the reassurance of knowing a human can step in, and that this reassurance is itself part of what makes a system feel safe and trustworthy. This suggests that AI is not yet at a place where it can replace humans, and that a human-in-the-loop remains necessary, especially in sensitive applications such as health and wellbeing, where consumers are disclosing personal, emotional, or high-stakes information and the risks of an inappropriate response are higher. In a health or wellbeing context, for example, a consumer left without that kind of support when something goes wrong may disengage from seeking help altogether, rather than push through to find a human alternative. Additionally, this risk is even more likely to affect those who already face barriers, such as language, digital literacy, or distrust of institutions. These consumers may be the ones most likely to be left with little support, and least able to find another way through when it fails them.
8. This is consistent with other evidence on how consumers feel about human oversight. Consumers generally do not approach AI as a replacement for humans, but rather expect human oversight to remain available and accessible as needed throughout their interactions with it. In mental health and wellbeing contexts specifically, stakeholders consistently identify robust human oversight, informed consent, and inclusive design as non-negotiable requirements, with concerns about privacy, bias, and inappropriate use remaining active areas of concern in current deployment.¹⁴ This pattern extends into other sectors as well, for example, as presented in a survey of over 500 consumers in the context of contact centre customer service found that nearly half still strongly prefer engaging with a human agent when given the choice, with only 5% expressing a preference for

¹² Ge, X., Xu, C., Misaki, D., Markus, H. R., & Tsai, J. L. (2024). How Culture Shapes What People Want From AI. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Article 95. <https://doi.org/10.1145/3613904.3642660>

¹³ Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2025). Sycophantic AI decreases prosocial intentions and promotes dependence. arXiv preprint arXiv:2510.01395. <https://doi.org/10.48550/arXiv.2510.01395>

¹⁴ Schoene, A. M., Mestre, R., Middleton, S. E., Lapedriza, A., Bernard, A., Lewis, M., Lardizabal, J., Movaghar, A., Fenelon, K. J., Lincoln, A. K., Gaverich, S., Scalise, D. II, Kumar, G. V., & Leeman, R. F. (2026). Responsible AI in mental healthcare: policy directions and stakeholder insights. *Frontiers in Public Health*, 14, 1814039. <https://doi.org/10.3389/fpubh.2026.1814039>

AI alone, suggesting that even as consumers use AI, they expect a human to remain reachable when needed.¹⁵

Question 14. How do existing terms and conditions include or reflect risks of AI?

9. Participants described terms and conditions as documents that nobody reads, citing their length and complexity as the primary barrier. One participant suggested they should be presented as bullet points with icons at a more accessible reading level. Participants felt strongly that holding users accountable for harms they could not have anticipated was not fair, given how inaccessible these documents are. Responsibility was seen as sitting with providers and developers rather than consumers, as one participant put it: *“you can’t think of someone who just doesn’t know and it might cause them harm and then maybe hold them accountable for it.”* This points to a deeper tension between the legal language companies use in these documents and the accessible language consumers actually need. Companies may have reasons for using precise legal language, such as regulatory compliance or limiting liability, but this has long been viewed negatively by consumers, and our participants experienced it the same way, as a barrier rather than a protection. Consumers may already be using AI to help interpret these documents, but for a group like the one we worked with, who already face additional barriers to understanding written English, this comes with its own risk. If someone cannot understand the original document well enough to check it, they also cannot check whether the AI’s interpretation of it was accurate.
10. This is consistent with the wider literature. Terms and conditions documents are rarely transparent or accessible. A longitudinal analysis of over 50,000 privacy policies found that while policies became longer following new regulations, but not clearer. They grew more vague, described data practices in unspecified terms, and offered users little meaningful choice over how their data was used.¹⁶ This pattern is especially concerning in the context of AI, where research has shown that consumers rarely read privacy policies and when they do they often misunderstand them. Even simplified policies have been shown to have close to no effect on consumer understanding of how their data will be used. The literature also suggests that the problem is not only how risks are presented but that privacy policies often do not include the information consumers need to understand how their data will be processed.¹⁷ A proposed solution discussed in the research above is to require privacy policies to be fully comprehensive rather than concise, and to use AI to help consumers navigate that length. While early testing found that AI was able to answer user questions about privacy policies accurately when written in this fuller format, suggesting AI could help close the gap between legal accuracy and consumer accessibility. But this testing was not done with a group like the one we worked with. It does not resolve the concern raised above, that a consumer who cannot fully understand the original document also cannot verify whether an AI’s interpretation of it is accurate.

Question 16. Who do consumers expect to be accountable when AI causes harm (provider, developer, user)? Does this differ for generative vs agentic AI?

11. Research shows that when AI causes harm, the nearest human actor, such as a frontline clinician, safety driver, or remote operator, often absorbs the moral blame for a wider system failure, even

¹⁵ Balto. (2025). 2025 Consumer Expectations Report: AI, Omnichannel, and the New Landscape of Consumer Expectations. <https://www.balto.ai/resources/research/2025-consumer-expectations-report-ai-omnichannel-and-the-new-landscape-of-consumer-expectations/>

¹⁶ Wagner, I. (2023). Privacy Policies Across the Ages: Content of Privacy Policies 1996–2021. *ACM Transactions on Privacy and Security*, 26(3), Article 32. <https://doi.org/10.1145/3590152>

¹⁷ Palka, P., Lippi, M., Lagioia, F., Liepiņa, R., & Sartor, G. (2024). No More Trade-Offs: GPT and Fully Informative Privacy Policies. arXiv:2402.00013. <https://doi.org/10.48550/arXiv.2402.00013>

when they had limited control or knowledge of the situation. This is referred to in the literature as the “moral crumple zone”. It describes a social effect, where the human closest to the failure is judged and blamed by others, regardless of how much control they actually had. This blame does not stay informal. It often translates into formal accountability as well, referred to in the literature as the “liability sink problem”, where responsibility falls on this same nearest human actor simply because they are the easiest person to hold legally responsible. These individuals are often not the most responsible for what went wrong, as they sit at the end of a long chain of design, development, and deployment decisions made by many others. Accountability should instead be distributed across all actors involved in the AI lifecycle, including developers, manufacturers, providers, and deployers, rather than concentrated at the point of consumer use.¹⁸

12. During our co-production workshops, participants shared their perspectives on AI and identified teenagers, non-English speakers, older people, and people facing digital divides and little to no AI literacy as the groups most at risk. Participants noted that these groups often do not recognise the risks they are exposed to, and expressed concern that these risks will grow as AI becomes more widely adopted. One participant raised concerns that people assume they are engaging in a private exchange when using AI, without realising that their information is being shared with third parties. She argued that people who lack this awareness cannot fairly be held responsible for what happens to their data. She drew on an everyday analogy, suggesting that AI risk communication should follow the model of public health guidance, clear, accessible, and proactively issued, much like the instructions distributed during a heatwave. She called for a straightforward set of guidelines informing people how to use AI responsibly, what to discuss, what to avoid, and what to share or withhold and risks associated with not following guidance.
13. Understanding who should be held accountable also depends on hearing directly from the people AI is most likely to affect. The Arab women we worked with shared that the workshops gave them space to engage with AI and its risks more broadly, and they felt empowered to explore them more deeply as a result, offering researchers insights drawn directly from their own lived experience. For many, it was the first time they had engaged with these issues in this way. Engagement of this kind does more than surface individual views rather it brings these voices to the fore and gives other stakeholders a way to understand cultural and societal experiences that are rarely captured in full through lab-based or experimental studies alone. This approach is not limited to the community we worked with. The same kind of engagement is needed with other underserved communities, whose experiences of AI are similarly unlikely to be captured through conventional research methods alone. The Arab women we worked with called for more dedicated AI literacy, critical engagement, and awareness opportunities, so that their voices can be heard, they are represented in how AI is developed and evaluated, and AI can be made genuinely safe for them, on their own terms.

Response Contributors

Lead Authors

- Dr Maria Waheed, University of Nottingham
- Dr Aisha Abubakar, University of Nottingham
- Dr Emily Thorn, University of Nottingham

¹⁸ Porter, Z., Ryan, P., Morgan, P., Al-Qaddoumi, J., Twomey, B., Noordhof, P., McDermid, J., & Habli, I. (2025). Unravelling Responsibility for AI. *Journal of Responsible Technology*, 23, 100124. <https://doi.org/10.1016/j.jrt.2025.100124>

- Dr Gisela Reyes-Cruz, University of Nottingham
- Dr Sarah Kiden, University of Southampton
- Sania Waheed, University of Southampton
- Sahar Waheed, National University of Sciences & Technology
- Marwa Soliman, Empowerment for HEYA
- Dr Aislinn Gomez Bergin, University of Nottingham

Please send any queries and comments about this response to maria.waheed@nottingham.ac.uk.

About AdSoLve

AdSoLve is a keystone project funded by Responsible AI UK. It brings together researchers from Queen Mary University of London, the University of Nottingham, the University of Sheffield, and the University of Warwick, working alongside partners including the National Health Service (NHS), Medicines and Healthcare products Regulatory Agency (MHRA), and National Institute for Health and Care Excellence (NICE). The project aims to create multi-faceted evaluation, including benchmarks with novel criteria, metrics and tasks as well as reference-free evaluation, for assessing the limitations of LLMs in real world settings, particularly in medicine and law, taking a responsible and ethical approach.

About Responsible AI UK

Responsible AI UK brings together researchers from across the four nations of the UK to understand how we should shape the development of AI to benefit people, communities and society. It is an open, multidisciplinary network, drawing on a wide range of academic disciplines. This stems from our conviction that developing responsible AI will require as much focus on the human, and human societies, as it does on AI. Funded by the Technology Missions Fund, we convene researchers, industry professionals, policymakers, and civil society organisations.

About the Voices of Safety Project

Voices of Safety is a research project funded through the SEED@Researcher Academy scheme at the University of Nottingham. The project develops community-defined requirements for safe and responsible AI-mediated wellbeing advice, grounded in the lived experiences of minority women from non-WEIRD backgrounds. It is piloted in collaboration with *Empowerment for Heya*, a Nottingham-based community of Arab women, positioning participants as experts of their own safety and wellbeing in order to address the gap between platform-defined AI safety and how safety, cultural appropriateness, and support are experienced within minoritised communities.