HACID - Deliverable

Guidelines for participatory AI

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101070588. UK Research and Innovation (UKRI) funds the Nesta and Met Office contributions to the HACID project.

Deliverable number:	D5.2
Due date:	30.08.2025
Nature[1]:	R
Dissemination Level[2]:	PU
Work Package:	WP5
Lead Beneficiary:	Nesta
Contributing Beneficiaries:	ISTC - CNR

[1] The following codes are admitted:
- R: Document, report (excluding the periodic and final reports)
- DEM: Demonstrator, pilot, prototype, plan designs
- DEC: Websites, patents filing, press & media actions, videos, etc.
- DATA: Data sets, microdata, etc.
- DMP: Data management plan
- ETHICS: Deliverables related to ethics issues.
- SECURITY: Deliverables related to security issues
- OTHER: Software, technical diagram, algorithms, models, etc.

[2] The following codes are admitted:
- PU – Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page)
- SEN – Sensitive, limited under the conditions of the Grant Agreement
- Classified R-UE/EU-R – EU RESTRICTED under the Commission Decision No2015/444
- Classified C-UE/EU-C – EU CONFIDENTIAL under the Commission Decision No2015/444
- Classified S-UE/EU-S – EU SECRET under the Commission Decision No2015/444

Document History

Version	Date	Description	Author	Partner
0.1	19.08.2025	Creation	Aleks Berditchevskaia	Nesta
0.2	26.08.2025	First draft	Aleks Berditchevskaia, Alexandra Albert, Christopher Edgar, Ewa Dominiak, Rita Marques	Nesta
0.3	27.08.2025	Partner Review	Vito Trianni	ISTC - CNR
0.4	28.08.2025-29.08.2025	Revisions	Aleks Berditchevskaia, Alexandra Albert	Nesta
0.5	29.08.2025	Final Review	Peter Baeck	Nesta

Design Guidelines for Participatory AI

1. Introduction

Participatory Artificial Intelligence (Participatory AI) refers to the intentional involvement of stakeholders - especially those directly affected by AI systems - in their design, evaluation, and deployment.[3] It draws on traditions from participatory design, human-centered computing, and democratic theory, aiming to ensure that AI tools reflect diverse values, meet real-world needs, and promote just outcomes.[4] Existing case studies using participatory AI tend to report at least some of the following benefits:

- It helps to improve model performance;
- It increases the usability, appropriateness and uptake of a tool for a given problem;
- It helps to align the tool with diverse values, needs and preferences;
- It can assist with anticipating and mitigating broader impacts, risks or harms; and
- It increases trust in AI by different stakeholders and trustworthiness of tools.

The rationale for participatory approaches in AI is clear: as AI systems increasingly mediate decisions across critical domains - healthcare, education, social services - decisions about how they function and for whom cannot be left solely to developers or technical experts. Participatory AI helps address power imbalances by giving voice to stakeholders who are often excluded, especially in sectors where the stakes are high and the consequences of poor implementation can be severe.[5]

Notable examples illustrate the value of this approach. In healthcare, participatory methods have helped redesign clinical risk algorithms by integrating feedback from frontline providers and patients, leading to tools that better reflect local values and concerns.[6] In crisis response, participatory AI has supported the co-design of decision-support tools that incorporate community knowledge and enhance trust and uptake in vulnerable regions.[7]

Despite these advances, much of the current AI development landscape remains limited in scope, struggling to mainstream the use of Participatory AI methods. Most participatory efforts focus on getting stakeholder feedback on incremental design changes or new

---

[3] https://www.nesta.org.uk/project/participatory-ai/. Accessed 20/08/2025.
[4] Berditchevskaia, A., Peach, K., and Malliaraki, E. (2021). Participatory AI for humanitarian innovation: a briefing paper. London: Nesta.
[5] Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2023, October). The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-23).
[6] Sendak, M. *et al.* (2020) '"The human body is a black box": supporting clinical decision-making with deep learning', in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. *FAT* '20: Conference on Fairness, Accountability, and Transparency*, Barcelona Spain: ACM, pp. 99–109. Available at: https://doi.org/10.1145/3351095.3372827.
[7] Berditchevskaia, A., Peach, K., Stewart, I., (2022). Localising AI for crisis response. London: Nesta.

features through A/B testing, or remain consultative in nature,[8] rather than more qualitative approaches that allow space for richer discussion or allow scope for challenging technical assumptions. This gap has led to calls for a more ambitious vision—one that not only includes stakeholders in AI design, but also empowers them to challenge assumptions.[9] The advent of foundation models and large language models (LLMs) in particular in recent years has posed further challenges to the participatory paradigm of technology development. The rise of these technologies has meant that AI-based solutions are increasingly powered by a small set of models that have not been built with datasets specific to downstream use-cases or domain expertise in mind, making upstream participatory interventions more challenging.[10] This has led to a greater focus on consultative approaches aiming to capture diverse public attitudes and preferences about use cases and governance, rather than a specific technology, led by global technology companies like Meta, independent research organisations such as the Collective Intelligence Project or public sector organisations[11,12,13]

Against this backdrop, the HACID project undertook a participatory design and research process to explore how clinicians and climate scientists perceive and evaluate the value of hybrid AI systems for decisions in their respective domains, focusing on the values that matter to them and the risks that concern them most. We undertook a mixed-methods approach, combining structured polling with qualitative discussion, aiming to surface practical insights that could be shared back with our partners on the HACID project to inform the active development of the HACID-DSS prototypes.

2. Aims

The HACID-DSS approach aims to support decision makers who are operating in high-stakes domains under circumstances of uncertainty. Our overarching purpose for developing the technology in a participatory way was to ensure the resulting tools are more trustworthy, aligned with end-user values, and ultimately practically useful to facilitate their uptake.

To address these aims, we designed and implemented participatory activities across the lifespan of the HACID project. We discussed and prioritised different options for participatory activities together with other partners in the consortium. We chose this approach in order to identify interventions that would be useful for the design of the prototypes, viable to deliver within the project's constraints and generate meaningful inputs that could be translated into technical decisions.

---

[8] Corbett, E., Denton, E., & Erete, S. (2023, October). Power and public participation in ai. In Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (pp. 1-13).
[9] Maas, J., & Inglés, A. M. (2024, October). Beyond Participatory AI. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (Vol. 7, pp. 932-942).
[10] Suresh, H. et al. (2024) 'Participation in the age of foundation models'. arXiv. Available at: http://arxiv.org/abs/2405.19479 (Accessed: 2 November 2024).
[11] https://globaldialogues.ai/ Accessed 20/08/2025
[12] https://www.bi.team/blogs/metas-community-forum-on-ai/. Accessed 20/08/2025
[13] https://sciencewise.org.uk/projects/ai-in-policing/. Accessed 20/08/2025

Key Performance Indicators (KPIs) 10-13 are the key metrics we established to measure our success at delivering a participatory AI approach to developing the HACID tools.

KPI number and description	Progress measure	Target	Project Objective	Lead Organisation
KPI-10 Process evaluation: How well do the evaluation methods capture the criteria that matter to decision makers?	Validation of the defined evaluation metrics by domain experts	Positive or very positive assessment by ≥ 80% of interviewed experts	**OBJ6: Evaluation:** develop an evaluation framework for decision support that covers diverse aspects relevant for the application context.	Nesta, CNR, MetO
KPI-11 Output evaluation: To what extent are the HACID-DSS outputs aligned with stakeholder values?	Assessment of the defined evaluation metrics by domain experts	Positive or very positive assessment by ≥ 80% of interviewed experts	**OBJ6: Evaluation:** develop an evaluation framework for decision support that covers diverse aspects relevant for the application context.	Nesta, MetO
KPI-12 Participatory design: Have we introduced new participatory approaches to AI development?	Number of participatory interventions throughout the tool development pipeline	≥ 5 participatory interventions	**OBJ7: Participatory AI:** deploy a participatory approach to the development of hybrid collective intelligence exploiting human expertise and AI.	Nesta and all other partners
KPI-13 Participatory evaluation: How well did we achieve the goals of participation?	Level of satisfaction reported by participants and relevant stakeholders.	High satisfaction and/or meaningful engagement reported by ≥ 80% of stakeholders.	**OBJ7: Participatory AI:** deploy a participatory approach to the development of hybrid collective intelligence exploiting human expertise and AI.	Nesta, CNR HDx, MetO

Table 1: KPIs 10-13 and how each KPI is measured. More detail can be found in D5.1

3. Participatory Activities

We developed a bespoke approach for each of the use cases, owing to the fact that the two DSS use cases were at different levels of maturity and stages of development. Specifically, in the medical diagnostics domain, we were adapting an existing tool with an established

community of users, however in the domain of climate services, we needed to first validate a concept and identify relevant decisions that would benefit from a HACID-like approach in order to build the tool from the ground up (including developing a knowledge graph).

Table 2 and 3 below provides an overview of the full set of participatory interventions tested across both use cases. See also Figures 1 and 9 which visualise these interventions against a simplified version of the HACID-DSS tool development pipeline.

Table 2: An overview of the full set of participatory interventions tested across the Medical Diagnostics use case

Medical Diagnostics HACID-DSS - participatory interventions		
Activity & Methods	**Purpose**	**Audience & recruitment method**
User research: semi-structured interviews	To understand the current challenges and opportunities in the clinical decision making process.	HDx staff; Primary and secondary care clinicians (convenience sampling through Human Dx platform) n=10
User research: focus-group (deliberation, survey - pair-based ranking)	To understand 1) what criteria clinicians prioritise for decision-making tools, 2) the benefit and limitations of current decision-making aids, 3) the appropriate scenarios for HACID and practical considerations for deployment.	Primary and secondary care clinicians (convenience sampling through Human Dx platform) n=6
Scenario-based risk assessment: Survey	To help identify and mitigate potential risks and maximise potential benefits for real life applications of a HACID-like tool from a “patient’s” perspective.	General public (purposive sampling through a research recruitment agency) n=241
Scenario-based risk assessment: deliberative-polling workshop	To help identify and mitigate potential risks and maximise potential benefits for real life applications of a HACID-like tool from an end-user perspective.	Healthcare professionals (purposive sampling through a research recruitment agency) n=8
Participatory evaluation: Deliberative workshop	To evaluate 1) how well the latest version of the prototype aligned with professionals’ values and 2) if our interventions achieved the stated goals of participation.	Primary and secondary care clinicians (purposive sampling through a research recruitment agency) n=12

Total interventions	5

Table 3: An overview of the full set of participatory interventions tested across the Climate Services use case

Climate Services HACID-DSS - participatory interventions		
Activity & Methods	**Purpose**	**Audience***
User research: semi-structured interviews	To identify user needs, challenges and goals for using climate data to make climate adaptation decisions.	Climate services clients e.g. National-level transport organisation, city-level transport organisation. n=5
User research: process mapping workshop and survey	To understand current processes and identify key challenges faced by climate scientists.	Climate services providers (Climate scientists and Met Office staff from different teams and services) n=12
Concept validation: feedback on lo-fi prototype	To validate the desirability, viability and feasibility of the concept.	Climate service stakeholders (academics, commercial climate services providers, policy audiences, Met Office staff) n=5
Knowledge graph interface design: Deliberative workshop, lo-fi prototyping	To understand preferences in terms of information discovery, access and visualisation.	Climate services providers n=6
Values elicitation: Deliberative workshop and values ranking	To involve domain experts in defining how information should be aggregated and to identify criteria the tool should be evaluated against.	Climate services providers n=5
Participatory evaluation: Deliberative workshop	To evaluate 1) how well the latest version of the prototype aligned with professionals' values and 2) if our interventions achieved the stated goals of participation.	Climate services providers (Met Office staff, different teams) n=7
	Total interventions	6

**Due to the specialised nature of the audience, we used convenience sampling through consortium partners for all activities in the climate services use case.*

For both use cases, we started with **user research** with the target professional communities to identify key design opportunities for a HACID-like technology and set the scope of the decision-making focus.

- Based on the findings from the user research activities, we proposed several viable design opportunities for the climate services HACID-DSS to the HACID consortium. One of these was prioritised for development (this is detailed in D7.1).
- Based on the findings from the user research activities, we proposed two viable prototypes for the medical diagnostics HACID-DSS, both of which extended the current functionality of the Human Dx tool: Cliniflow and Multi-Disciplinary Teams (these are explained in detail in D6.1)

We ended our participatory interventions for both use cases with a deliberative **participatory evaluation**.

We tested several other interventions in between. We used a two-pronged approach to deciding which interventions to implement during the project:

1) **Proactive**, consortium-led deliberation and prioritisation - after completing the user-research activities for both use cases, we held a workshop with consortium partners where we presented different options for participatory activities, alongside their purpose and resource requirements. Values elicitation and Participatory evaluation activities were prioritised using this approach.

2) **Reactive** workshops/interventions - based on emerging needs from technical partners (e.g. Interface design workshop) or due to constraints in partners' availability to contribute to participatory interventions (e.g. scenario-based risk assessment).

Parallel to its development, the HACID tool was integrated into a commercially developed product (the HDx platform/app), which meant partners had to balance research needs with product priorities — including user experience, development timelines, and reputational considerations (e.g., avoiding being perceived as a research or data-gathering tool). At times, it was challenging to navigate these tensions and we had to reprioritise and replan participatory activities several times as a result.

3.1 Overview of a participatory framework adapted to Use Case 1: Medical diagnostics

Figure 1 shows the framework for participatory activities for the medical diagnostics use case. These included user research, a scenario-based risk assessment and participatory evaluation.

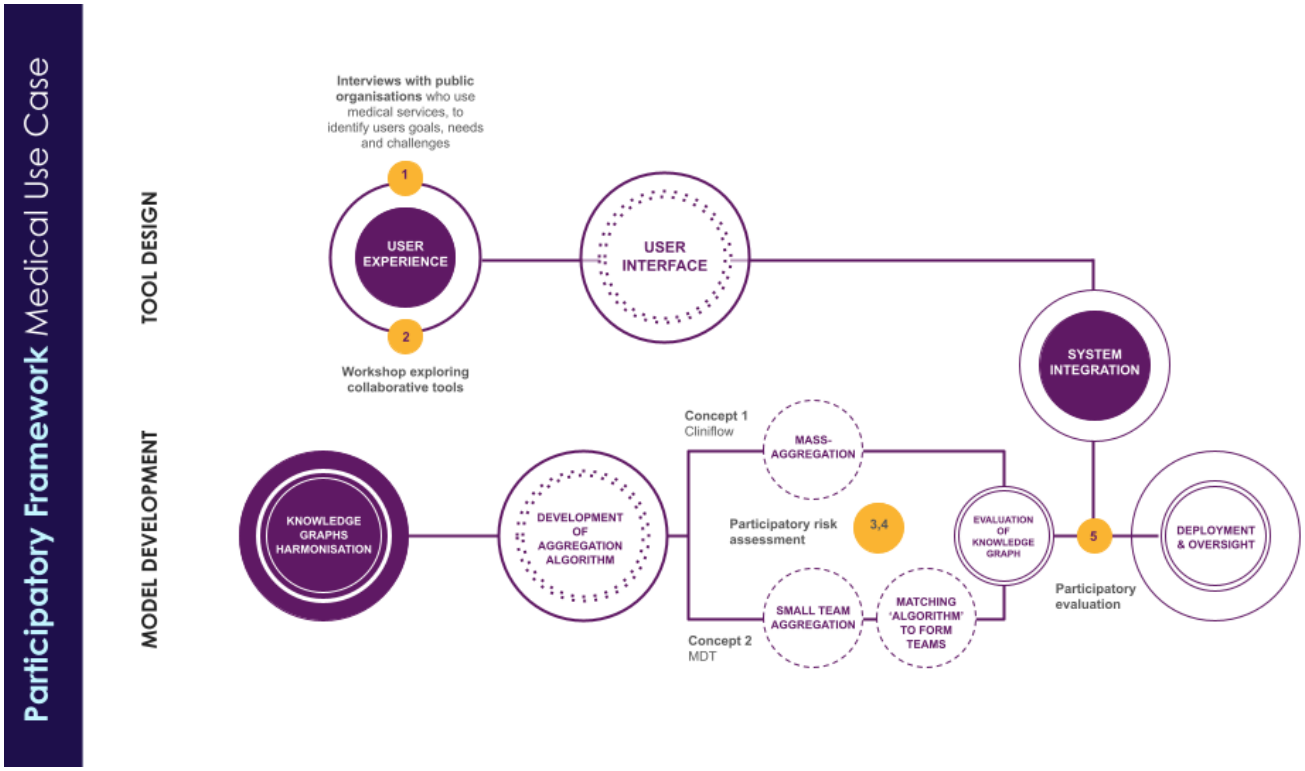


Figure 1: The participatory AI framework adapted to use case 1: Medical diagnostics

4.1.1 User research for Use Case 1: Medical Diagnostics

A complete description of the methodological approaches, key findings and design outputs from user research activities were documented in D6.1 Requirements for HACID-DSS in medical diagnostics. We provide a concise overview of the activities below and direct interested readers to D6.1 for a detailed overview.

The user research phase for the medical diagnostics HACID use case employed semi-structured interviews and a collaborative workshop. The primary objectives were to understand how a decision support system (DSS) could be used in medical diagnostics, identify key user features, and validate the system's value proposition of using a hybrid intelligence approach to support real-time decision making.

Our core research questions focused on understanding:

- the current clinical decision-making process,
- the factors that build trust in decision support tools, and,
- the potential value of collective intelligence in increasing confidence in outputs.

The participants in this phase of research included the technical teams developing the Human Dx platform, as well as clinicians from different geographical locations across Europe, the United States, South America, and Asia, working across primary and secondary

care, and representing various specialisms and seniority levels. This diversity was intended to provide a broad foundational understanding of the user landscape. These interviews informed the development of two distinct personas that represent different user journeys and a distinct set of challenges in diagnostic decision making: the General Practitioner and the Specialist (Fig xx).

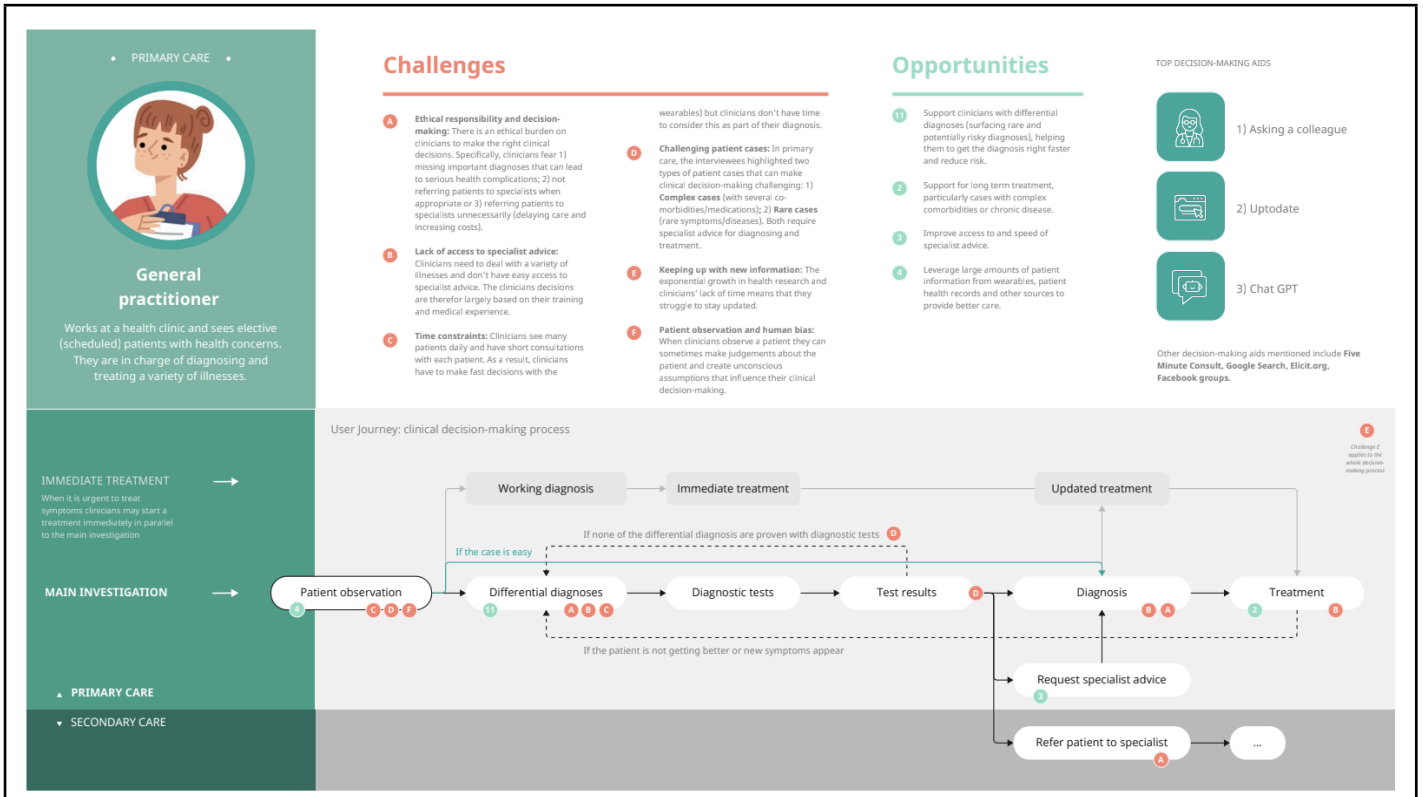


Figure 2: Example persona and user journey: General practitioner, who were characterised by a reliance on informal and relatively unstructured information sources, leading to the development of the Cliniflow concept.

The research also revealed that clinicians' top criteria or values for assessing the robustness of a decision support tool are 1) Quantifying uncertainty through clear probability/confidence ratings, 2) Accuracy and 3) Accountability for errors. These findings directly influenced the design principles for the HACID-DSS, which emphasised trustworthiness, patient-awareness, the need to avoid human bias, and time efficiency.

These research outputs were then used to develop two design concepts: "Cliniflow" and "Virtual Multi-Disciplinary Team Collaboration". Our design strategy was to create two distinct concepts that cater to the specific needs of the two personas and the differences in their decision-making workflows. The General Practitioner's need for a fast, on-demand tool aligns with the concept of Cliniflow, a hybrid intelligence tool providing immediate, real-time AI-based support. Conversely, the Specialist's need for inter-disciplinary collaboration was addressed by a concept we called the Virtual Multi-Disciplinary Team (MDT), which facilitates structured, asynchronous deliberation among a pre-selected group of experts.

A key limitation of the user research was the relatively small sample size of 13 clinicians due to recruitment challenges. This small scope may not fully represent the diverse perspectives within the larger clinician community. We triangulated these early-stage findings through subsequent participatory activities and experiments designed to test the Cliniflow and MDT workflows (see deliverables produced in WP4).

3.1.2 Risk assessment of Use Case 1: Medical Diagnostics

Overview of the activities

This PAI intervention consisted of two discrete activities:

1. A survey for the general public
2. A deliberative workshop with clinicians.

Both activities required participants to engage with three tangible clinical scenarios (using engaging and accessible visual assets) for using the HACID tool in diagnostic decision making, and to identify their top concerns about risks in these different cases.

The workshop brought together a small group of eight NHS clinicians from a range of clinical backgrounds, representing a mixture of roles across primary and secondary care. It was held online to accommodate the limited availability of professionals. Delivered via a bespoke digital platform,[14] the workshop combined audio case studies, interactive polling, and facilitated discussion to surface expert insights into the real-world implications of AI-supported clinical tools.

In addition to this, adding a novel layer to this participatory activity, the workshop featured the presentation of survey data from the general public, detailing their perceptions of risk to the same scenarios, and our clinicians were asked to reflect on their perceptions of risk and think about how abstract risks translate into real ethical and professional dilemmas - and where existing governance may fall short.

What Do We Mean by Risk, and Why Does it Matter?

As artificial intelligence becomes more embedded in clinical decision-making, the conversation is shifting from whether it poses risks to how we identify, assess, and address them. These risks—ranging from algorithmic bias and data privacy breaches to reduced transparency and unclear accountability—are real, but they also present opportunities to build more trustworthy, equitable systems.[15] In particular, hybrid AI models, which integrate human expertise with machine intelligence, are emerging as a promising approach. By

[14] https://zg-app.com/
[15] Siafakas, N., & Vasarmidi, E. (2024). Risks of Artificial Intelligence (AI) in Medicine. Pneumon, 37(3).

preserving clinician oversight while enhancing diagnostic consistency and coverage, these systems not only help mitigate many common concerns but also open space for more context-sensitive and participatory design—ensuring that risk management in AI is not just reactive, but proactive and human-centred.[16,17,18]

To ground our exploration of risk, we drew heavily on recent work which has has synthesised years of research investigating risks in healthcare research,[19] which has developed a useful framework for understanding the current state of risk in AI-enhanced healthcare. Among the most urgent concerns are: algorithmic bias, which can marginalise underrepresented populations; opacity in system reasoning, which undermines clinician trust; data privacy vulnerabilities; and accountability gaps when errors occur in hybrid human-AI decision pathways. These risks are not hypothetical - they have tangible impacts on patient care, public confidence, and clinical workflow.

Reinforcing the concerns of academic literature, institutions like the Organisation for Economic Co-operation and Development (OECD) and the European Parliament have echoed these concerns. The OECD's 2024 report on AI in healthcare stresses that risks stem not only from flawed models, but from how they are governed, explained, and integrated into already strained health systems.[20] It warns of dangers like deepening inequity, fragmenting care, or failing to protect vulnerable populations. Likewise, the EU's Science and Technology Options Assessment (STOA) Panel identifies seven major clusters of risk, including patient harm, data misuse, lack of transparency, and regulatory mismatch with fast-evolving AI capabilities.[21] Importantly, many of these risks are relational - they emerge not from the technology alone, but from the interplay between AI and the humans, institutions, and norms that surround it.

Yet, in spite of these risks, surveys of public and professional opinion to AI in healthcare paint a nuanced picture. Insights from two recent large-scale surveys conducted by the Health Foundation show that while public and NHS staff attitudes towards AI were broadly supportive - 54% of the public and 76% of NHS staff supported AI use in patient care - many still expressed discomfort with how AI might affect relational care and decision-making processes. For example, 53% of the public said AI would make them feel more distant from

[16] Kurvers, R. H., Nuzzolese, A. G., Russo, A., Barabucci, G., Herzog, S. M., & Trianni, V. (2023). Automating hybrid collective intelligence in open-ended medical diagnostics. Proceedings of the National Academy of Sciences, 120(34), e2221473120.
[17] Zöller, N., Berger, J., Lin, I., Fu, N., Komarneni, J., Barabucci, G., ... & Herzog, S. M. (2024). Human-AI collectives produce the most accurate differential diagnoses. arXiv preprint arXiv:2406.14981.
[18] Bajwa, J., Munir, U., Nori, A., & Williams, B. (2021). Artificial intelligence in healthcare: transforming the practice of medicine. *Future healthcare journal*, *8*(2), e188-e194.
[19] Chustecki M. Benefits and Risks of AI in Health Care: Narrative Review. Interact J Med Res. 2024 Nov 18;13:e53616. doi: 10.2196/53616. PMID: 39556817; PMCID: PMC11612599.
[20] OECD. (2024). AI in health: Huge potential, huge risks. OECD Publishing. https://www.oecd.org/health/ai-in-health-report
[21] European Parliamentary Research Service. (2022). Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts (STOA Study). European Union. https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2022)729512

NHS staff, while 65% of NHS workers feared it would reduce their connection with patients.[22] Support also fell sharply when AI outputs were not subject to human review, highlighting a strong desire to keep humans in the loop.

In the HACID Risk Assessment Workshop, clinicians engaged with AI-enabled clinical scenarios not just as evaluators, but as stakeholders and co-designers. Their reflections helped surface how abstract risks translate into real ethical and professional dilemmas, and where existing governance may fall short. Building on prior research, we extended this work by exploring how these concerns play out in the context of hybrid AI systems like HACID. In doing so, we move toward an understanding of risk that is not only technically sound, but socially grounded and professionally credible.

The risk assessment took place in two parts: part 1 entailed a survey of the public's perceptions of risks, and part 2 entailed a workshop with clinicians where we ran several activities: a review of clinical scenarios, a risk prioritisation activity and a responsibility rating activity.

Part 1: Public Risk Perception Survey

To generate stimuli for the workshop discussions, a public-facing online survey was conducted in advance of the workshop to explore how members of the UK public perceive risks associated with hybrid AI in healthcare. A nationally representative sample of 241 UK adults was presented with **three narrated clinical scenarios**,[23] which consisted of three audio-narrated case studies to illustrate distinct applications of the HACID system in diagnostic support (See Appendix for a detailed description):

- Rare disease diagnosis – highlighting the potential for AI to flag overlooked conditions using limited clinical data.
- Mental health diagnosis – exploring how hybrid AI might support clinical judgement in contexts of diagnostic ambiguity.
- Hormonal imbalances and patient-reported data – examining personalised diagnosis through patient-contributed information and demographic profiling.

For each scenario, respondents were asked to identify up to three concerns via open-text responses. To generate the stimuli of a ranked list of the general public's perceived risks, a topic modelling pipeline was applied to a subset of early responses (n=241[24]). See Appendix for a full overview of the sample and methodology for thematic coding. Figure 3 shows an example of the resulting themes for the 'Rare disease diagnosis' scenario. The results from the public survey, top risks for each individual scenario and an aggregated list, were used as

[22] Health Foundation. (2024). *AI in health care: What do the public and NHS staff think?* https://www.health.org.uk/reports-and-analysis/analysis/ai-in-health-care-what-do-the-public-and-nhs-staff-think

[23] Each respondent was presented with only 2 scenarios in order to not exceed recommended survey good practice and sustain attention.

[24] While 286 public responses were ultimately collected, analysis for the workshop was based on an initial subset of 241 submissions, due to internal deadlines and resourcing constraints. All responses were stored, and the additional data may be included in a later analysis.

prompts for comparing professional and public perceptions of AI risk and trust during the workshop with clinicians.

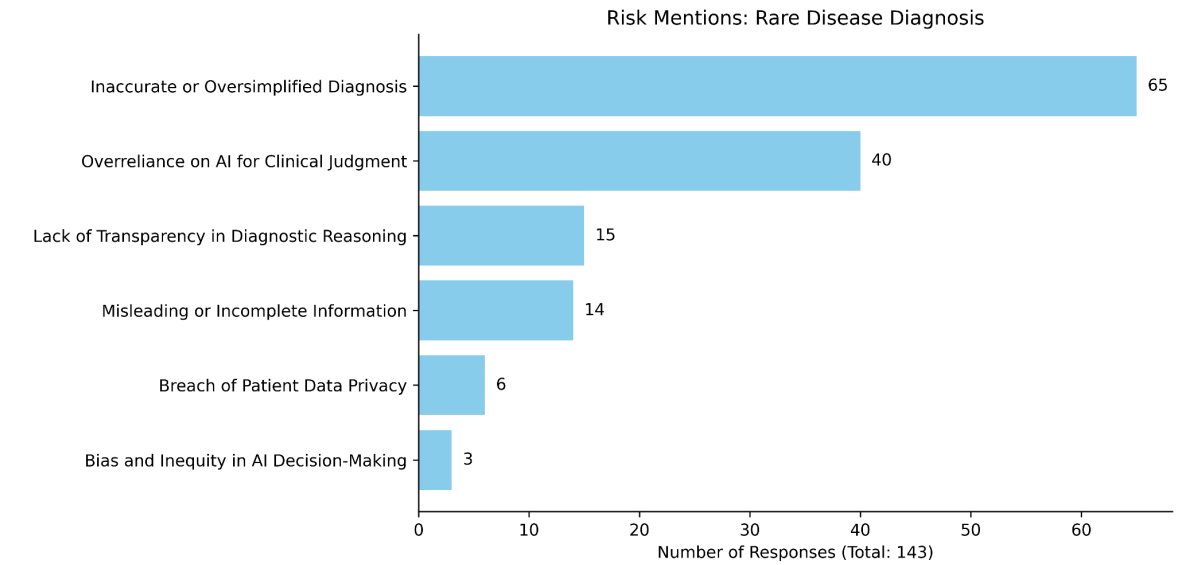


Figure 3: Example stimuli illustrating general public's top concerns relating to the use of hybrid AI in healthcare. Presented during the assessment of the Rare Disease Diagnosis scenario (n=143, only a subset of the full sample were presented with this specific scenario).

Part 2: Clinicians Workshop

The workshop opened with a short animated video introducing the aims of the workshop and the concept of hybrid AI in clinical contexts. Participants then completed **a short survey** covering demographics, professional background, familiarity with AI, and their initial views on the personal and societal impacts of AI.

Participants were then invited to discuss the same three clinical scenarios as used in the public survey.

Following each scenario, participants engaged in structured, facilitated discussion. They were invited to share initial reflections, identify perceived risks, and evaluate how the AI-supported approach compared to standard clinical practice.

To deepen the dialogue, participants were shown visualisations of risk-related concerns derived from the public survey responses for each scenario. Facilitators used targeted prompts to explore whether participants felt the HACID system introduced improvements, new risks, or shifted boundaries of clinical responsibility.

Participants were then asked to **rank six core risks associated with AI in healthcare**, based on the public survey themes and related expert literature. After submitting their initial

rankings, they viewed aggregated public risk rankings across all three scenarios and were invited to reflect on and discuss any similarities or differences. Participants could re-rank the risks based on new information or perspectives raised in the discussion, and were also given the opportunity to suggest any missing risks.

In the final section of the workshop, participants **rated the degree of responsibility** that different stakeholder groups (such as, AI developers, clinicians, hospital leadership, regulators) should bear for ensuring the safe and ethical use of AI in healthcare. A brief facilitated discussion followed, focusing on perceived gaps or tensions in current governance arrangements.

The workshop concluded with a brief evaluation phase designed to evaluate different aspects of the participatory process[25] and any shifts in participants' attitudes towards AI.

Results: Key Findings

Scenario 1 - Rare Disease Diagnosis: clinicians expressed a nuanced view of HACID's potential. Several participants saw clear value in using a hybrid-AI tool to accelerate diagnosis for complex or rare conditions that traditionally take years to identify. One participant framed this as a key benefit, noting that patients often "wait years and years for a diagnosis," while another supported this by suggesting HACID could assist in rapidly synthesising relevant literature, streamlining the initial stages of clinical reasoning. However, concerns quickly surfaced around the risk of over-investigation, with a participant warning that presenting an expanded list of potential diagnoses could burden clinicians and create psychological stress for patients. Additional concerns included:

- **Transparency:** Difficulty in communicating the AI's role in the diagnostic process to patients.
- **Accountability:** Uncertainty over legal responsibility if the AI contributes to a misdiagnosis.
- **Data security:** Worries about including patient cases in AI training databases, given past data breaches.
- **Clinician deskilling:** A risk that clinicians may become overly reliant on HACID, reducing their diagnostic independence.

One participant reflected that HACID appeared safer than many current AI tools, so long as it was used in conjunction with clinical judgement. The discussion underscored the balance between reducing diagnostic delays and introducing new ethical and procedural burdens.

> *"This solution I see as a little bit safer. In this case, we are also relying on your knowledge, and that makes it different and therefore safer. I am not sure about the current AI solutions."*

[25] Participants completed a five-item survey assessing dimensions of clarity, inclusion, and value. These items were adapted from Nesta's Participatory AI for Humanitarian Innovation framework and mirrored those used in the HACID Values Elicitation Workshop. They were intentionally shortened and simplified to minimise participant burden and align with the digital delivery format.

Scenario 2 - mental health diagnosis: participants expressed a broad consensus that HACID would have limited applicability for mental health diagnosis. Clinicians noted key differences in mental health practice:

- Diagnosis often takes a backseat to understanding individual needs
- Many patients present with multiple past diagnoses, making classification less useful
- The clinical focus is often relational, rather than diagnostic

That said, some felt HACID could still play a secondary role as a tool for information gathering or second opinions. However, concerns were raised about using it in sensitive cases—particularly with individuals experiencing paranoia or psychosis—where the involvement of AI might cause distress or exacerbate symptoms.

"We are working from a slightly different angle. We tend to understand what is happening with the person rather than place a diagnosis. We work with people who have received 10 different diagnoses throughout their life. There is less clinical utility here."

"One condition that would be really sensitive is paranoia and psychosis. Especially if the patient is experiencing paranoia about the technology. In this case I can see how it could be particularly dangerous and difficult to explain to the patient how we arrived at the diagnosis and used this tool."

Indicative quotes from workshop participants

Participants also considered how AI might impact the clinician-patient relationship. Concerns included:

- **Undermining authority:** Patients may feel empowered to self-diagnose if AI is seen as the primary decision-maker.
- **Public misunderstanding:** Surprise was expressed that "oversimplified or inaccurate diagnosis" ranked highly among public concerns, possibly reflecting a gap in understanding how clinical judgement typically works in mental health.
- **Condition-specific safety:** HACID might be inappropriate for some conditions, those suffering from paranoid delusions may be distressed by the use of AI technology, if they are already wary of it.

"Some people already may self-diagnose. We have a lot of people coming to the clinic prepared with the use of the internet and questioning our diagnosis. In this term HACID may enhance it even more…"

"Patients might not feel comfortable coming back to you again if they have particular thoughts about the use of technology and it was less accurate."

Indicative quotes from workshop participants

The discussion reinforced the idea that HACID, as currently imagined, does not align well with evolving mental health practice, which is shifting away from rigid diagnosis and towards holistic, person-centred care.

Scenario 3 - Hormonal Health Imbalance: Participants saw both potential benefits and risks in applying HACID to hormonal health diagnoses. Key concerns included cost and time - HACID could produce long lists of differential diagnoses due to overlapping symptoms, potentially straining resources. Another concern was around overdiagnosis and the risk of unnecessary investigations arising from overly broad diagnostic suggestions.

On the other hand, participants acknowledged the potential for HACID to speed up diagnosis for conditions that are often delayed or missed—particularly in women's health (e.g., endometriosis). Still, there was no clear consensus on whether HACID would improve safety or outcomes compared to existing tools.

> *"Pros: A lot of women's conditions take a long time to be diagnosed by GPs. Example being endometriosis. The AI may expedite the diagnosis.*
>
> *Cons: I agree with the other speaker — the cost. Also, the risk of misdiagnosing conditions."*
>
> ***Indicative quotes from workshop participants***

Risk prioritisation and alignment between public and clinicians' concerns

When comparing their own ranked list of risks with those generated by the general public, participants generally found the two to be broadly aligned.[26] Participants ranked overreliance on AI for Clinical Judgment consistently as the most concerning risk. Inaccurate or oversimplified diagnosis also increased in perceived importance from pre-discussion to post-discussion. However, concerns about "Breach of Patient Data Privacy" and "Lack of Transparency in Diagnostic Reasoning" declined slightly in importance by the end of the workshop.

These results suggest that discussing the risks and reflecting on how their rankings compared to those of the general public, had little impact on participants' overall prioritisation of risks for hybrid AI in healthcare. However, it did appear to reinforce existing views: participants became more concerned about the risks they were already worried about, and less concerned about those they had previously deprioritised. This suggests that deliberation had a consolidating effect on their perspectives.

Building on this, participants identified several additional risks they felt were missing or under-emphasised, both in discussion and through free-text responses.

[26] As noted, participants were shown public risk rankings prior to completing their initial ranking (i.e. after each scenario). When asked, *"To what extent did public concerns influence your ranking?"* (1 = Not at all, 5 = A great deal), most reported little to no influence. Responses ranged from 1 to 3, with a mean of 1.63, suggesting that public views had minimal impact on their individual rankings.

A key theme was equity in care, particularly around uneven access to AI tools across better- and worse-funded services. Participants also raised **concerns about patient confidentiality, data quality and sources, and legal and accountability issues**, particularly around unclear responsibility for AI-supported decisions. The risk of workforce displacement—AI replacing or undermining clinicians—was mentioned, along with fears of over-reliance on AI leading to a decline in clinical judgement:

Further risks included incorrect treatment decisions, variable digital literacy among clinicians, and challenges in communicating AI decisions to patients, which some linked to broader concerns about trust and transparency. Participants noted that older patients, and sometimes older clinicians, may struggle more with trust or understanding of AI tools.

Finally, several participants emphasised gaps in upskilling support for clinicians adapting to such tools, questioning whether current training infrastructure was adequate. Some also speculated that public concerns about "inaccurate diagnosis" may reflect limited understanding of the often opaque nature of clinical reasoning itself, rather than specific distrust in AI.

"My concern would be in areas where there is more or less funding or private care. Are we going to see equity in implementing such a solution?"

"Over-reliance on clinical judgment. It looks like this won't replace that but I'm sure there might be some clinicians who would allow it to, potentially over time."

"Do we need to keep training doctors if we're spending too much money just to make a diagnosis?"

Indicative quotes from workshop participants

Responsibility Rating for safe use of AI

On the question of responsibility for ensuring AI tools like HACID are safe and ethical, participants agreed that accountability should be shared, with the greatest burden falling on developers and clinicians. Developers were expected to ensure the tool's performance and safety, while clinicians were viewed as ultimately responsible for how such tools are applied in real-world settings.

"…AI developers are responsible for ensuring that the literature and the outcomes are accurate. Clinicians are the ones making decisions when using the tool, so if you use the tool, the clinician takes responsibility. These days we have a lot of influence from pharmaceutical companies. Advocacy groups may play a smaller role in responsibility, as they can be more susceptible to external influences."

These positions are reflected in the accompanying polling data (see Figure 4), where AI developers and individual clinicians received the highest responsibility ratings from participants. By contrast, groups such as patient advocacy organisations were generally seen as having a more limited role. This was mirrored in the discussion, where several participants voiced scepticism about the independence of these groups, particularly when their activities may be shaped by commercial interests. Existing clinical protocols were seen

as a useful foundation for managing shared responsibility, while advisory panels and similar actors were perceived as playing a supporting role.

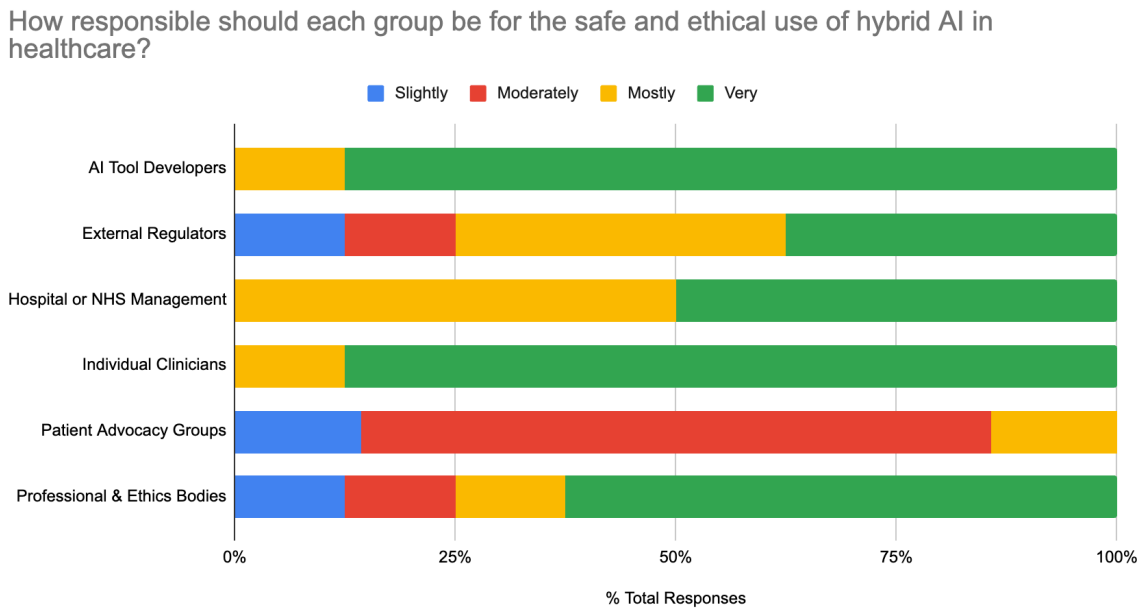


Figure 4: Ratings of actor responsibility for the safe, ethical use of hybrid AI in healthcare.

Participants also reflected on the **practical challenges of implementing tools like HACID**. Experiences varied widely. Some reported smooth integration of similar tools in outpatient settings, particularly when well supported by training and clear clinical pathways. Others, especially those in hospital environments, highlighted the scale and complexity of change management, including system interoperability and cultural resistance.

To succeed, participants felt any implementation would need to be gradual and carefully managed, with open communication and transparency. This would also need to be backed by robust training, tailored to clinicians' varying levels of technical confidence, and be sensitive to generational differences, particularly among older clinicians who may be more hesitant to adopt new technologies.

"We have recently launched a chatbot that comes with a provisional diagnosis. It's meant for us as clinicians. We've been able to integrate it into our services, and it seems like HACID could be a smooth transition as well."

"In a hospital environment, it would be difficult to launch something like this. It would take time and effort for the change to happen."

“It should be launched slowly and carefully, with a lot of sharing and openness. Careful management would be key.”

Indicative quotes from workshop participants

Conclusion

This workshop provided critical insight into how clinicians perceive the risks, responsibilities, and practical implications of hybrid AI in diagnostic decision-making. While participants generally agreed with public concerns - particularly around overreliance on AI and inaccurate or oversimplified diagnoses - they also surfaced additional, profession-specific risks rooted in clinical realities, such as legal accountability, workforce impacts, and inequities in access to AI tools.

Participants emphasised that the perceived value and safety of hybrid AI tools like HACID depend heavily on context. The tool was seen as potentially beneficial for conditions marked by diagnostic delay, such as rare diseases and hormonal health issues, but less appropriate in domains like mental health, where relational care and diagnostic ambiguity are central.

In total, the following risk areas were identified, either through structured rankings or qualitative input:

- Overreliance on AI, leading to potential deskilling of clinicians
- Inaccurate or oversimplified diagnoses
- Misleading or incomplete information
- Bias and inequity in AI decision-making
- Breach of patient data privacy
- Lack of transparency in AI reasoning
- Legal and accountability issues surrounding AI-supported decisions
- Equity of access to AI tools across different regions and service types
- Workforce displacement, including replacement or undermining of clinicians
- Incorrect treatment decisions, particularly if AI outputs are accepted uncritically
- Patient trust and communication challenges, especially in vulnerable groups
- Digital literacy gaps among clinicians

These risks reinforce that safe implementation of AI requires more than technical functionality. Participants called for clear accountability frameworks, robust training and support, and careful integration within existing workflows. The process also demonstrated that participatory engagement with clinicians not only surfaces practical risks but strengthens buy-in and trust—key factors for the successful adoption of AI in healthcare.

Evaluating the Participatory Process

We developed a brief evaluation survey (see Appendix for survey items) for participants to reflect on the quality and structure of the participatory process, aligned with KPI-13. The survey was developed to align with the recommendations for good practice outlined in

Nesta's *Participatory AI* framework[27] and Delgado et al's *Participatory turn in AI Design.*[28] We used these survey items consistently across the different activities.

Participants rated various aspects of the participatory process highly. As shown in Figure 5 clarity of purpose and ease of participation received particularly strong ratings. The item *"I was able to express what I think is important for trustworthy AI"* received the lowest ranking of the different items.

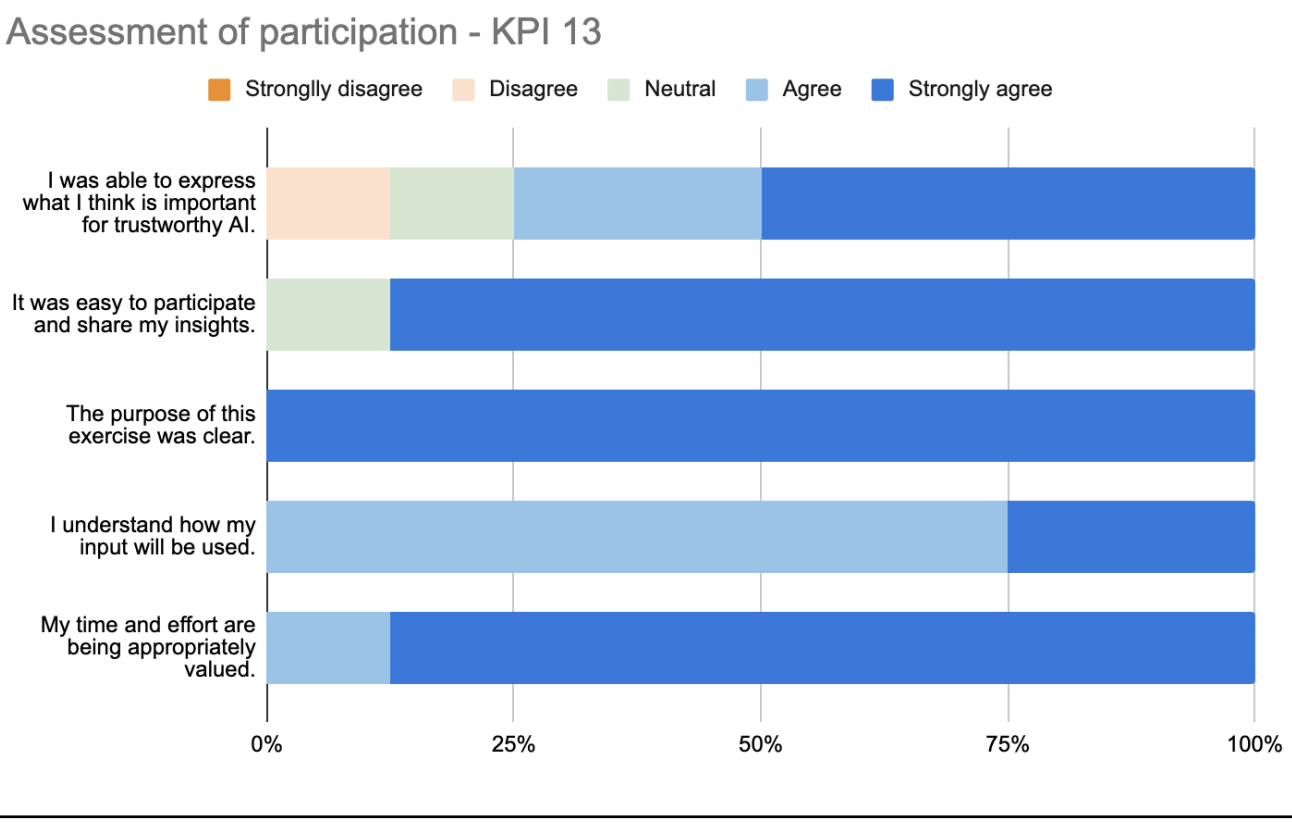


Figure 5: Evaluation of the participatory process. (n=8)

Five Key Takeaways from the HACID Risk Assessment Workshop

1. **Involving clinicians in participatory design adds value and improves engagement**
   Participants rated the workshop process highly, particularly in terms of clarity,

[27] Berditchevskaia, A., Peach, K., and Malliaraki, E. (2021). Participatory AI for humanitarian innovation: a briefing paper. London: Nesta.
[28] Delgado, F. et al. (2023) 'The participatory turn in AI design: Theoretical foundations and the current state of practice', in Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. New York, NY, USA: Association for Computing Machinery (EAAMO '23), pp. 1–23. Available at: https://doi.org/10.1145/3617694.3623261.

relevance, and ease of participation. This suggests that participatory approaches are an effective way to surface insights, build engagement, and foster clinician trust in emerging AI tools.

2. **Clinician perceptions of risk broadly aligned with those of the public**
   Polling data showed that clinicians, like members of the public, viewed overreliance on AI and inaccurate or oversimplified diagnosis as the most concerning risks. This alignment may support the development of shared communication and governance strategies.
3. **Clinicians surfaced additional risks not captured in public data**
   While overall rankings were similar, clinicians identified additional concerns including equity of access, workforce displacement, legal accountability, and the potential erosion of clinical judgement—issues shaped by their practical experience and professional responsibilities.
4. **Context and clinical setting shape the perceived value of hybrid AI**
   Perceptions of HACID varied by clinical context. While it was seen as valuable in areas like rare disease and hormonal health (where diagnostic delays are common), it was perceived as less applicable in mental health, where diagnosis is less central and relational care is prioritised.
5. **Responsibility for AI safety is seen as shared—but not equal**
   Participants viewed AI developers and clinicians as bearing the greatest responsibility for the safe and ethical use of hybrid AI. This was reflected in both qualitative discussions and quantitative polling. In contrast, groups such as patient advocacy organisations were viewed with some scepticism, particularly where commercial influence was seen as potentially compromising their independence.

3.1.3 Participatory Evaluation of Use Case 1: Medical Diagnostics

Objectives

The main objective for this final part of the design process of the HACID tool was to complete a participatory evaluation of the tool together with clinicians responsible for primary and secondary care. Within that we sought to understand clinicians' perspectives on how well the tool aligns with certain professional standards or values, and to understand the contributing factors to clinicians' preferences for the different types of information presented in the tool (combination between LLM-generated and expert-generated advice) and the different stages at which such information is presented to contributors (before or after they contribute their response to a case).

Methodology

We worked with a research recruitment agency to engage a participant group of 12 clinicians from a diverse mix of backgrounds. They spanned primary and secondary care, with a mix of seniority levels of specialisms and levels of experience, from a range of ethnicities, ages, and geographic regions. Participants were given an incentive fee of £120 for participation in the workshop, as well as an additional smaller incentive (£30) to undertake the pre-workshop task.

Our approach consisted of 3 key activities:

1. **A short pre-workshop task (15mins)** which entailed downloading, registering and accessing the Human Dx app to familiarise themselves with its functionality. We asked participants to record initial reflections on:
    a. the ease of finding relevant insights, and to note down in the response form any further questions they may have had.
    b. how a tool like Human Dx might fit into their daily clinical workflow, including scenarios where it might be most useful.
2. A **90min deliberative workshop** to surface participants' views on what is meant by AI and share experiences with each other, including how AI is being used in healthcare. We also introduced the participatory development process of the Human Dx tool, what the tool is meant for and how it works.
    a. Activity 1 - a participatory assessment of Human Dx based on existing values,
    b. Activity 2 - a deliberation to refine how information is presented in the Human Dx tool.
3. **A post-workshop feedback survey** was administered to participants to evaluate the quality of the activities along five dimensions directly related to good participatory design.

Results

The pre-workshop task, where clinicians engaged in free exploration of the Human Dx workflows, revealed that a majority of participants were overall positive about the app (Figure 6), and its potential value for clinical decision making. (see Appendix for a summary of individual reflections from the pre-task)

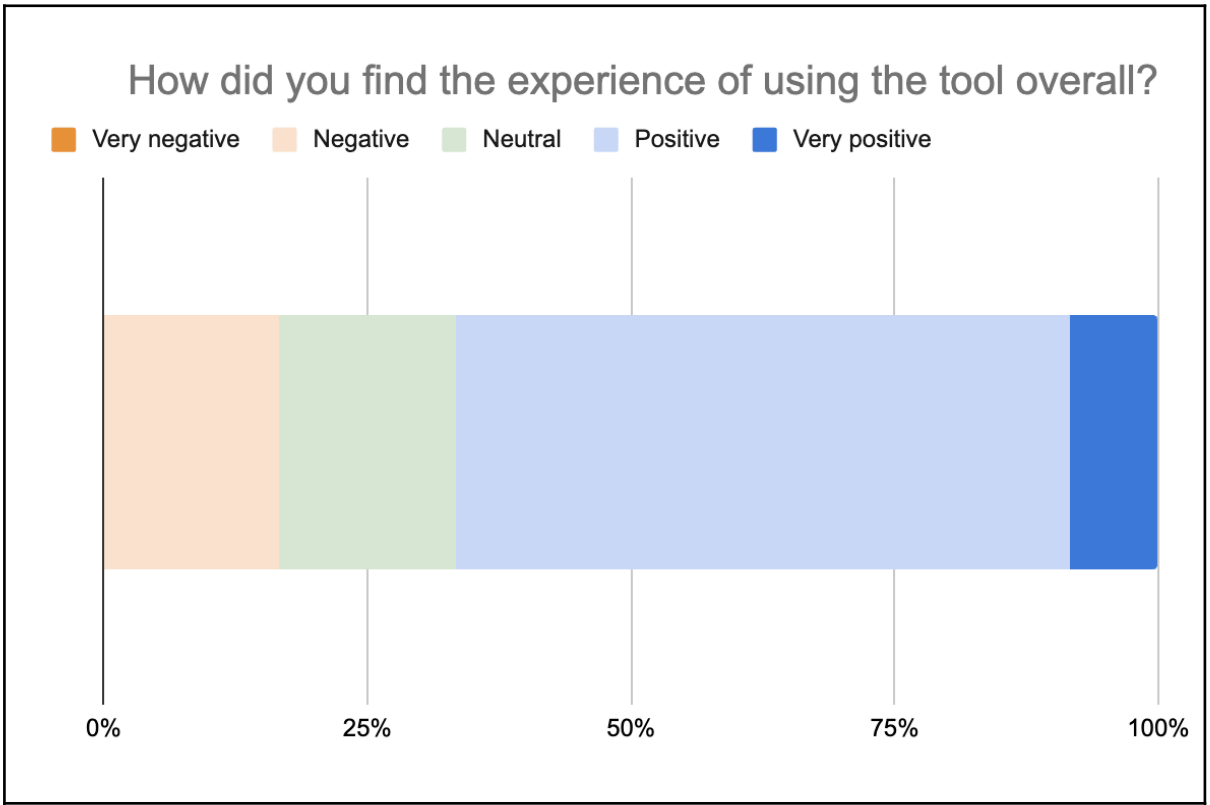


Figure 6: Participants' assessment of the tool during the workshop pre-task.

Activity 1 - Assessing how well the HACID tool aligns with clinician's values

The first evaluation activity focused on assessing how well the HACID prototype aligns with the values considered important by clinicians. As a starting point, we used the prioritised list of features for trustworthy decision support systems suggested by clinicians during the user research activities in the early phases of the project (Figure 7). We first verified whether these values were considered as important by the workshop participants. The clinicians were asked whether anything was missing from the list and whether they agreed with the ranking. There was broad agreement with the ranking/prioritisation, although it was recognised that all of the features on the list could be important in certain contexts.

Figure 7: A prioritised list of features for trustworthy decision support systems suggested by clinicians during the user research activities

'I definitely agree with those top three really because I suppose that's what we are encouraged to do as healthcare professionals.'

'As a pharmacist, if something goes wrong with a clinical decision that I make, as long as I can back up my findings with a reference or source it's ok. But with AI, on a lot of the apps and the information that I read, you just get a spam of text telling you what the answer is with no reference to that. And if there was some sort of reference to where the information came from, then I might feel more convinced of the accuracy and confident to even look at it by myself to confirm I'm happy with the suggestions it makes.'

Indicative responses from workshop participants

After the validation of the prioritised values, we moved on to the assessment of the HACID prototype against the top ranking values. Participants were given a simple poll to assess: *which value does the Human Dx tool align with particularly well?* Two values received the highest count (see Table 4): **Avoiding human bias** and **Efficiency**.

Question	Answer	Count
Of the values presented here, which is the Human Dx tool **particularly well** aligned with?	Provides error / confidence rating	2
	Accuracy	1
	Accountability / liability for errors	1
	Aggregates patient data	0

	Avoids human bias	**4**
	Efficiency	**4**
	Aligned with ethics	0
Of the values presented here, which is the Human Dx tool **most poorly** aligned with?	Provides error / confidence rating	3
	Accuracy	0
	Accountability / liability for errors	**5**
	Aggregates patient data	1
	Avoids human bias	0
	Efficiency	0
	Aligned with ethics	3

Table 4: Poll Results from Participatory Evaluation of Medical Diagnostics Use Case

When asked to unpack their responses, participants reflected:

'With avoiding human bias. I think in some ways that's a good thing but the nuances of working in mental health and having human experience which comes into play with our decision-making means that sometimes we need to have conversations with other clinicians and have a think about what emotes us as people as well and reflect on that.'

'I voted for efficiency. I think because although I'm pretty certain that it should never actually replace your clinical judgment, it could save a lot of time in terms of looking through guidelines, looking through departmental guidelines, online things, BMJ, NICE guidelines, that sort of thing to make sure that you're definitely following the right path. It'd be a lot quicker than manually going and looking all of those things up.'

Indicative responses from workshop participants

In the next poll, participants assessed: *which value does the Human Dx tool align with least well?* Respondents selected **Accountability or liability for errors.** During subsequent deliberation, they explained their selection. They recognised that liability would be an issue for both the AI-generated and crowdsourced advice, particularly if it wasn't possible to verify the identity of professionals providing responses.

'If I discuss a difficult case with a consultant at the hospital, I will usually document that in the notes. It's very difficult to be able to at this stage anyway put 'discuss with Human Dx' in the case notes. And then if there is an issue down the line, how do you come back to the app? I don't know how much responsibility Human Dx is going to be taking for that decision.'

‘Everybody needs to be accountable for the decisions that they make especially with patient care and I don't think AI would be accountable to anyone. It does not have any regulating framework. It does not have any regulating body like the GMC for doctors - it would just be a gray area.’

Indicative responses from workshop participants

Activity 2: Part 1 - Assessing the balance between LLM-generated and clinician-generated advice

In the final deliberation, participants were asked to think about the balance between AI generated advice and crowdsourced advice from other clinicians and their preferences for when and how different types of information might be presented if they’re posting a contribution. We explored **3 new scenarios** for how the app could provide advice to clinicians’ queries:

- solely advice from LLMs
- advice crowdsourced from other experts
- advice from both LLMs and experts

For each of the scenarios, we used facilitated deliberation to explore the benefits and concerns the participants anticipated.

Overall, participants’ didn’t state a fixed preference for the appropriate balance between LLM and expert generated advice. Participants suggested that the optimal mix would be to provide both AI and crowdsourced opinions, especially for non-urgent cases, allowing users to review similar past cases for information rather than waiting for immediate answers.

The clinicians highlighted several concerns about the verifiability and trustworthiness of crowdsourced contributions. They also expressed concerns that crowdsourced responses were slow and impractical compared to calling a local specialist. In general, the participants preferred the idea of having multiple LLMs generate responses but they wanted to understand more about the way the responses were being combined or ranked between the different models.

Participants recognised that optimising the combination of human and LLM advice could help to address the limitations of either approach on its own. For example, they suggested using AI to filter inappropriate or misleading human responses, and using verified professionals to review and correct misleading information generated by LLMs. They also emphasised the added-value of the app in providing sufficient information to avoid omissions, rather than substituting clinical judgment or providing definitive diagnoses, recognising that the clinician remains responsible for decision-making.

‘When it comes to getting the clinician results or the actual human results versus the AI results, it's about the accountability and reliability of those… I think that would be worrying that you're getting people that actually are not qualified to give a medical opinion on a

situation or they give an opinion and that gets embedded within your AI. Is there a way that the AI can have AI to pick up human responses that are actually completely inappropriate or misleading?'

'I don't mind the idea of just using LLMs as much…but I would like that you could inform it of which guidelines to be using and referencing for.'

'I would say actually I would feel more comfortable with it being taken from multiple AI sources just to see where they agree and where they disagree…just to give you a little bit more confidence that they haven't made an obvious mistake because presumably they wouldn't all make the same mistake at the same time.'

'I suppose the best thing would be to give both the opinions. It does take some time to get the crowdsourced information which could be the issue but I suppose if it's a non-urgent case or something that's been playing on your mind for a while and if you get notified about a response or something a bit like Reddit where you can look into what's happened or see what the response is so far.'

Indicative responses from workshop participants

Activity 2: Part 2 - Assessing the timing of different types of advice

In the final part of the discussion, participants were asked about the timing of the presentation of different types of information in the Human Dx app. The purpose of the discussion here was to understand what factors contribute to clinicians' preferences in terms of the timing of LLMs' input into the diagnostic process. The advice and summaries from LLMs can be presented before clinicians contribute their answers (pre-advice). Alternatively clinicians can post their answer/response and the advice from other experts and the LLMs summaries are presented after this (post-advice). Overall, participants strongly preferred posting their answers before seeing others' responses, to ensure independent thought and prevent copying or unconscious bias, which were considered important to maintain the quality of contributions. However, participants also recognised that reviewing other contributions or advice first could be beneficial for personal learning or add more detail and nuance to your own contribution after seeing others' responses.

'I would definitely say the second scenario [post-advice]. I think if you see all the other responses first, the unconscious mind is going to slip into, well, yeah, I agree with that, I'm just going to write the same thing. Whereas if you're just writing it without seeing anyone else first, you have to actually think and you have to actually put your own opinion.'

'From a learning point of view, being able to see in scenario one probably would be most efficient and quick from your own learning. But I suppose in terms of contribution to the overall platform then scenario two is probably best.'

'Being able to go back and change your answer after the fact might be a useful thing. If you noticed that the LLMs had given something incorrect, being able to say actually I don't think you should agree with that and this is why could provide useful context for whoever's asked the question in the first place.'

Indicative responses from workshop participants

Conclusion

Clinicians validated the values that informed our design principles for the HACID tool and were appropriate criteria for assessing the trustworthiness of the prototype. We learned that the Human Dx tool aligns most closely with stakeholder values around avoiding human bias, and efficiency, and aligns least well with the stakeholder value of accountability or liability for errors. The input gathered during Activity 2 of the workshop will be shared with partners as an input to refine the next version of the app, ensuring it aligns with clinicians' preferences and daily work realities. The results will also be used to help with interpreting the findings from the experiments planned as part of WP4, designed around different workflows for the Cliniflow and MDTprototypes.

Evaluating the Participatory Process

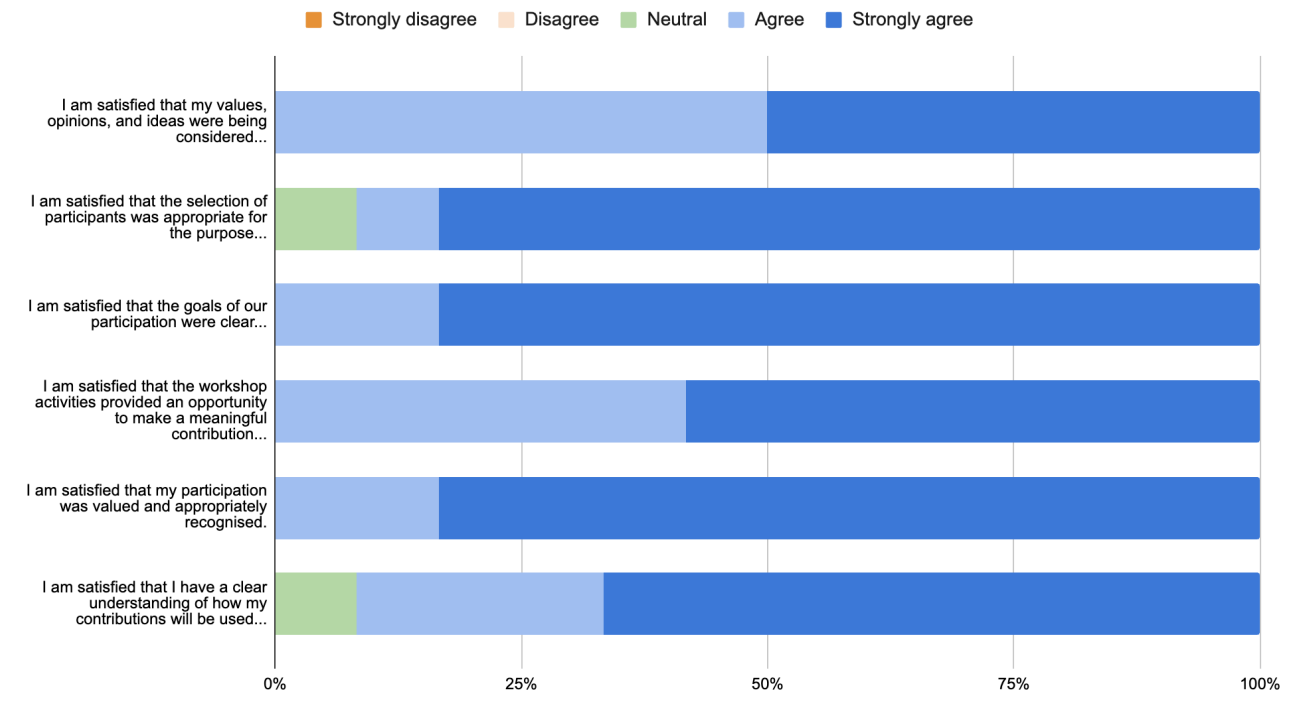


Figure 8: Final assessment of the participatory evaluation (medical diagnostics) by workshop participants.

(n=12) We met our KPI target across all dimensions. Likert statements have been shortened for visual clarity.

At the very end of the workshop, participants were asked to complete an online feedback survey to assess the participatory intervention. Participants rated all of the dimensions of participation highly. This was notable because the participatory evaluation of the medical diagnostics prototype was the final participatory intervention we held and showed an improvement across dimensions that had been rated less highly during previous participatory engagements ("the diversity of the participants engaged" and "understanding how their contributions will be used").

3.2 Overview of the participatory framework adapted to Use Case 2: Climate services

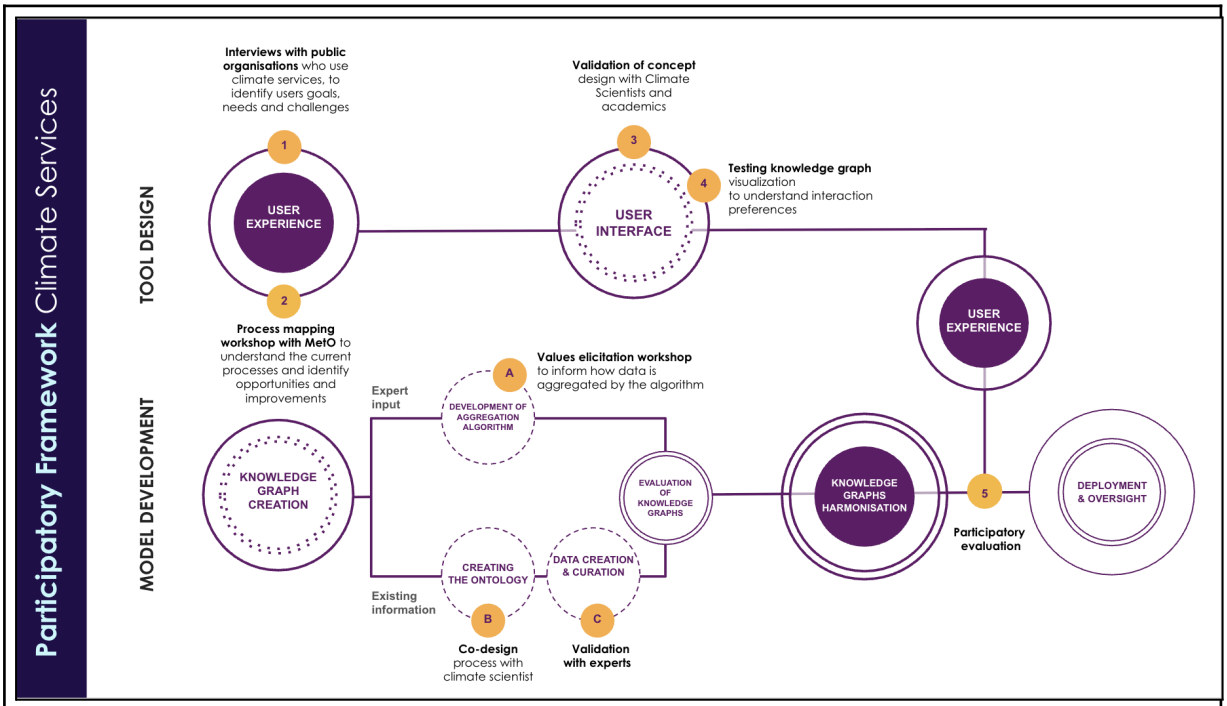


Figure 9: The participatory AI framework adapted to use case 2: Climate Services

4.2.1 User research for Use Case 2: Climate Services

A complete description of the methodological approaches, key findings and design outputs from user research activities were documented in D7.1 Requirements from climate services. We provide a concise overview of the activities below and direct interested readers to D7.1 for a detailed overview.

The research employed a mix of qualitative methods, including semi-structured interviews, surveys, and collaborative workshops to engage a total of 21 participants from two main stakeholder groups:

- **Climate services clients** that use climate services, such as a transport company. The research explored their motivations, needs, and challenges when using climate data for adaptation decisions.
- **Climate services providers** involved in delivering climate services. This group participated in a system-mapping workshop and a survey to uncover their processes and challenges.

The core research questions we focused on during user research explored:

- how clients of climate services make decisions and use climate information,
- The current process of delivering climate services and the challenges faced by climate scientists, and,
- the information needs, challenges, and constraints of climate change experts in data-driven decision-making.

The research findings were analyzed and grouped into four key themes that served as design opportunities for the HACID-DSS prototype:

- Understanding User Needs: User requirements for climate services are often broad and change over time, requiring customized, iterative solutions.
- Addressing Uncertainty: Climate information is inherently non-static and uncertain due to evolving models. Climate scientists struggle to provide a level of certainty they are comfortable with. Users also need to understand the level of uncertainty to make informed decisions.
- Granularity of Data: External organizations often require more granular data, but climate projections are limited by data availability and computational power. Combining projections with other datasets is often necessary to achieve the desired granularity.
- Information Navigation and Discovery: Users find it difficult to navigate the abundance of available climate datasets and resources. They often resort to paying consultants for information that is already publicly available but hard to find.

Based on criteria of feasibility, user value, and innovation, the project consortium prioritised the opportunity of **"Addressing the non-static/uncertain nature of climate information".** This led to the design of a prototype that supports climate scientists in making decisions on appropriate information sources, methodologies, and future scenarios by crowdsourcing solutions from other experts. The prototype aims to enable peer learning and create consensus, ultimately leading to more accurate, trusted, and efficient solutions. We mapped out the user journeys and a viable service delivery model in a prototype service blueprint (Figure 10). This blueprint has informed the development of the HACID-DSS for climate services.

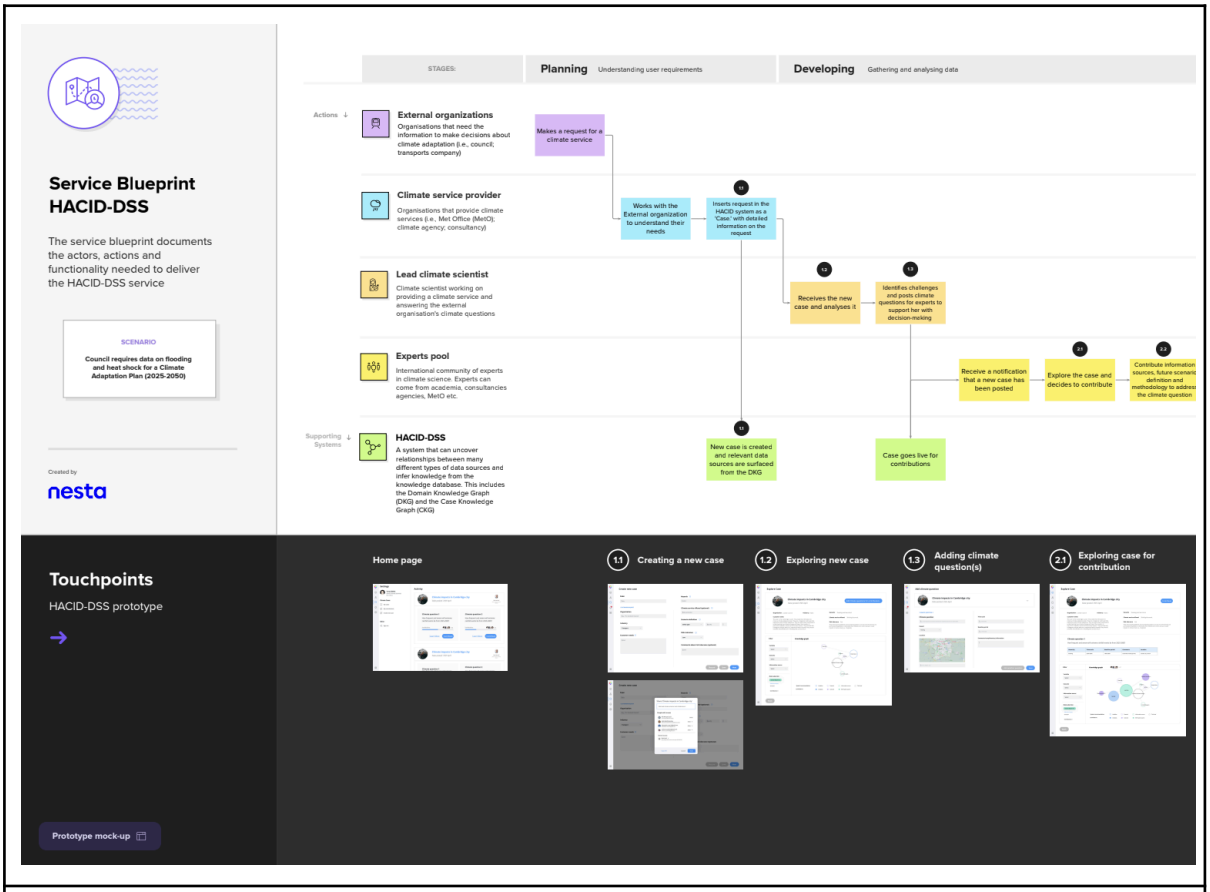


Figure 10: Detail of the service blueprint for the HACID-DSS climate services prototype.

3.2.2 Knowledge Graph Interface Design for Use Case 2: Climate Services

Workshop Overview and Objectives

We used a deliberative online workshop to bring together 6 climate scientists from institutions including the University of Oxford, the Met Office, and the Turing Institute. The goal was to collect data reflecting the user-interface preferences of two different potential end-users of the HACID tool: a contributor providing their solution to the submitted query (that the HACID system could draw upon when providing an answer to the query); or a climate scientist submitting a query to be answered by the system.

Methodology

Participants were introduced to the idea of hybrid artificial intelligence and the HACID project, with a particular emphasis on the use of its underlying knowledge graph technology. They were asked to consider three options for different ways of displaying the outputs of and interacting with the knowledge graph underpinning the system's answer to a given query:

- **Visual Query:** A branching diagram representing the links between different key attributes of the users' query, and related information. Typically allows users to expand attributes to view related information.
- **Concept Map:** A diagram that visually represents relationships between concepts and attributes related to the query and solution. Users are able to expand nodes and explore related sources of information.
- **Basic Chatbot:** An interactive chatbot, capable of answering user queries using natural language.

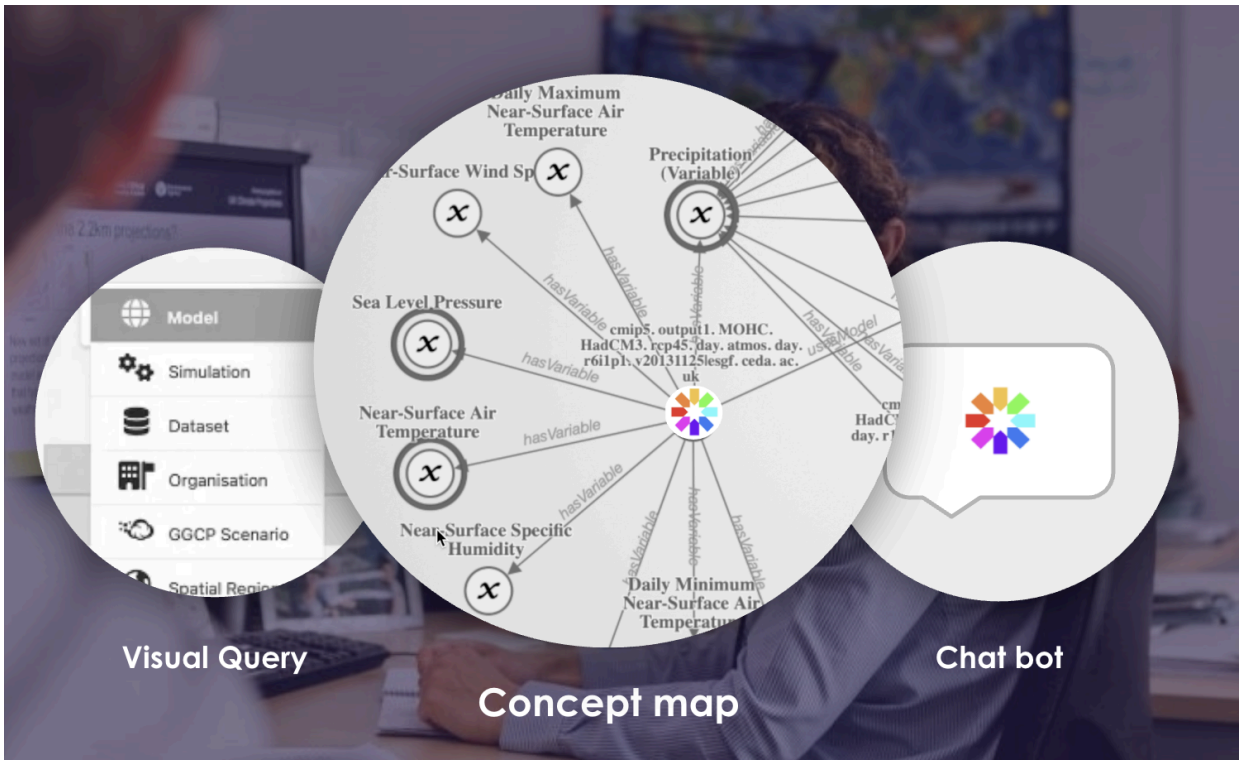


Figure 11: Visualisation of the 3 different interface options for interacting with the knowledge graph.

Participants were shown videos providing an overview of what each of these three options would look like (Figure 11), and how the interfaces would be navigated. Participants were then given two scenarios, and asked to rate each solution as either 'most useful', 'neutral', or 'least useful', followed by a short discussion to explore their preferences in more depth.

In scenario 1, participants were asked to respond to the question: "As a contributor, which prototype would you find most useful (comparatively)?"
In scenario 2, participants were asked to respond to the question: "As a lead climate scientist, which prototype would you find most useful (comparatively)?"

Results

For both scenarios, participants provided their answers to the polling questions, and then engaged in a short discussion to express related thoughts and opinions in more depth.

Points were totalled across participants for each scenario to provide a summary of the usefulness of each interface (see Figure 12).

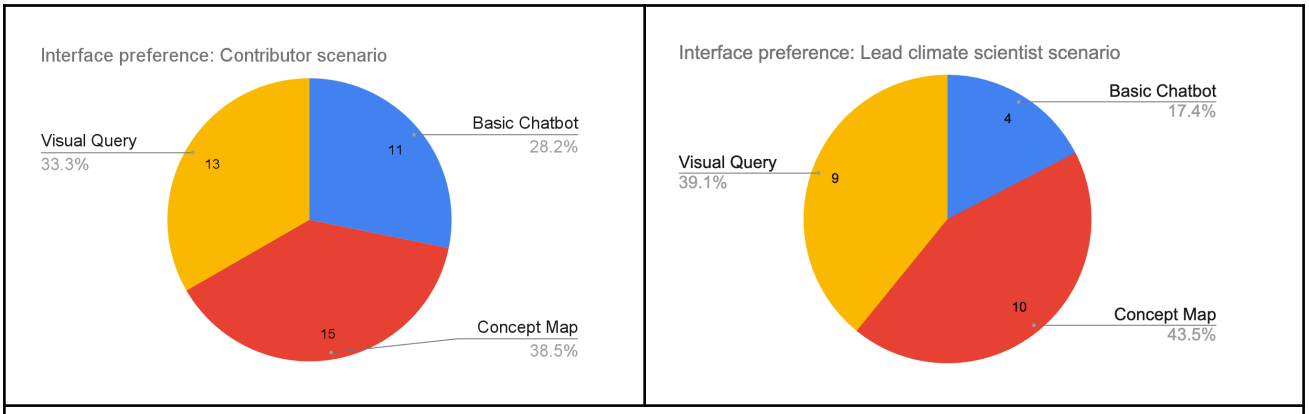


Figure 12. Polling data: Total assigned points for each interface option for both scenarios.

For both scenarios, the **concept map was the preferred interface**, with participants articulating three key reasons for their choice:

1. **Exploratory Use**: The concept map was seen as the best way to facilitate exploration within the knowledge graph, especially in scenarios where users are uncertain about what they are looking for or when dealing with dynamic and large datasets.

> *"An advantage that something like the concept map is going to have over a visual query, is that you might not find what you're looking for unless you can query sources - the concept map's gonna allow you to explore."*
>
> *"The value I could really get out of this would be using the concept map to make sure I've sort of covered everything because… just making sure I've sort of considered all avenues."*

2. **Visualisation of Gaps**: Participants believe that the concept map helps identify gaps in knowledge more effectively than other interfaces.

> "*Concept map allows us to visualise gaps in knowledge."*
> *"Finding holes in knowledge is very difficult to do, but visually probably more easy."*

3. **Intuitive Nature**: Concept maps are seen as more intuitive compared to visual queries, which might be more traditional but less intuitive for some users.

> *"It is quite intuitive. I think once you get into it and get into the rhythm of it."*
> *"If you're using it multiple times, then actually filtering it with your eyes or doing some sort of mental filtering, the more you use it, I guess the easier it gets."*

Participants also recognised the **potential value of the visual query** based interface. The **visual query** was recognised as providing a more traditional interaction that the participants were already comfortable with, which would facilitate uptake. They preferred this approach for cases where they already had a clear understanding of what they were looking for in the dataset

"If I know what I'm looking for - I'd use a visual query."

"The visual query is kind of a more traditional approach to narrowing down things and more familiar to people."

The chatbot interface **was the least preferred** due to scepticism about their reliability. However, the participants recognised that chatbots could be useful for more casual users who may not be familiar with the knowledge graph's structure.

"Phrasing the question in the right way is key to that. And that's not always easy to do."

"I'd be quite concerned about hallucinations."

"You don't have to know how the underlying info is structured, this makes it more appropriate for the more casual user."

Conclusion

These preferences were fed back to the technical team at CNR and informed the next phase of the prototype development, leading to a concept map with visual querying features.

3.2.3 Values Elicitation for Use Case 2: Climate Services

In the context of AI, *values* are the normative principles and evaluative criteria that guide what we consider "good," "acceptable," or "trustworthy" system behaviour.[29] [30] As AI systems

[29] Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International journal of human-computer studies, 146, 102551.
[30] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication, (2020-1).

are increasingly integrated into consequential decisions, ensuring that these systems align with values has become increasingly relevant.[31]

International bodies and industry consortia have converged on a shared vocabulary of values for "trustworthy AI." For instance, the OECD AI Principles emphasise human-centredness, transparency, robustness, and accountability[32], while the EU's Ethics Guidelines for Trustworthy AI outline requirements such as fairness, explainability, and societal well-being.[33] These frameworks provide important high-level guidance, but they leave open the question of how to translate abstract principles into practical system design choices.

This gap has driven increased interest in value-sensitive design and participatory AI methods, which aim to embed stakeholder values throughout the AI lifecycle.[34] By involving affected communities, domain experts, and end users in eliciting and reconciling the values that matter most, these approaches help tailor system behaviours to real-world contexts, boosting both legitimacy and trustworthiness.[35]

Crucially, values are neither monolithic nor static: they vary across domains, stakeholder roles, and the nature of the task.[36] Participatory processes can reveal not only which values participants endorse, but how they negotiate trade-offs among competing concerns - insights that are critical if AI systems are to earn genuine trust and deliver context-sensitive, socially robust outcomes.

Workshop Overview and Objectives

The HACID Values Elicitation workshop brought together 5 climate scientists from institutions including the University of Oxford, the Met Office, and the Turing Institute. The workshop was held online and aimed to employ participatory methods to understand what values shape climate experts' trust in AI-driven decision support systems (DSS), particularly the HACID platform.

Activity 1: Investigating What Impacts Trust in Individual Solutions

This activity aimed to elicit user values by examining how participants evaluated the trustworthiness of individual solutions to a climate services query. The exercise was designed to provide a granular, bottom-up understanding of what builds or reduces trust in system-generated outputs—without pre-defining values or ethical criteria.

---

[31] Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2023). Operationalising AI ethics: barriers, enablers and next steps. AI & SOCIETY, 1-13.
[32] Organisation for Economic Co-operation and Development (OECD). (2024). OECD AI principles. https://oecd.ai/en/ai-principles
[33] European Commission. (2019). Ethics guidelines for trustworthy AI. Publications Office of the European Union. https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1
[34] Berditchevskaia, A., Peach, K., and Malliaraki, E. (2021). Participatory AI for humanitarian innovation: a briefing paper. London: Nesta.
[35] Umbrello, S., & Van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. AI and Ethics, 1(3), 283-296.
[36] Klingefjord, O., Lowe, R., & Edelman, J. (2024). What are human values, and how do we align AI to them?. *arXiv preprint arXiv:2404.10636*.

Participants evaluated six different responses to a climate services query:

> *"What are the most appropriate sources of climate data and information that can be shared with engineers and asset managers to assess the risk of flooding and impacts of heat extremes over London for the 2050s?"*

Each response varied by source type (expert, AI-generated, or peer-reviewed article) and was structured across three sections:

- **Source** – e.g. institutional affiliation, role title, location, community endorsement score,[37]
- **Response** – the substantive climate data recommendations,
- **Rationale** – the justification provided for the recommendations.

For each solution, participants rated their overall trust (1–5 scale), selected up to three specific pieces of information that influenced their trust, and briefly explained why each piece mattered to them. To analyse these explanations, a thematic coding approach was applied independently by two researchers. Each highlighted item was classified as having either increased, decreased, or had no impact on trust, and then categorised under one or more of six emergent trust labels: **accuracy**, **community endorsement**, **credibility**, **detail**, **relevance**, and **transparency**. These labels captured the most salient factors driving trust-related judgments and underpin the results presented in the next section.

Results: Key Findings and Patterns

Participants consistently trusted solutions from known, reputable sources more than those with anonymous or AI-generated provenance. Solution 6 (a peer-reviewed article) and Solution 1 (a senior Met Office scientist) received the highest trust scores, while chatbot or unattributed responses received the lowest.

Analysis of participant justifications revealed that detail, credibility, and relevance were the most frequently cited reasons impacting trust (see Figure 13).

[37] This is a score derived from upvoting or other means of rating a response by other participants.

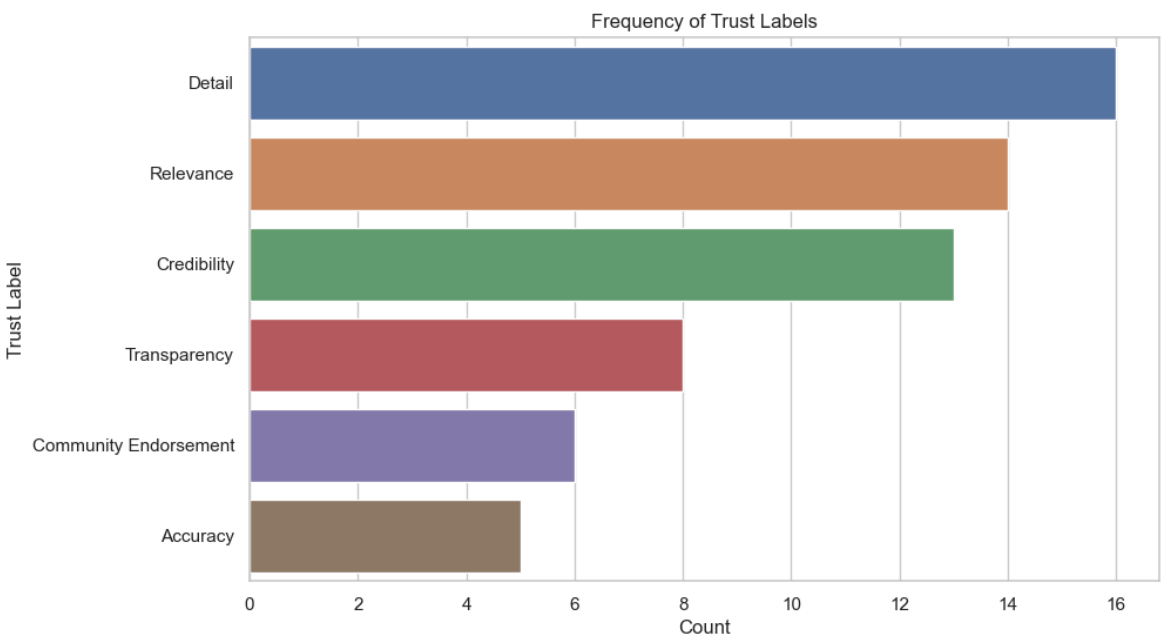


Figure 13: Frequency of trust label application across all participant responses

To explore this further, we distinguished between reasons that increased versus decreased trust. Information (i.e. participant supplied reasons) was coded as increasing trust 34 times and decreasing it 25 times.

- When trust increased, participants most often pointed to:
    - Rich detail or nuance in the response
    - Recognisable or credible sources
    - Strong relevance to the query

- When trust decreased, the dominant concerns were:
    - Lack of relevance
    - Opaque sourcing or missing provenance
    - Inaccurate or questionable information

These findings suggest that overall detail, source credibility, and relevance are central to trust in AI-supported climate services. However, positive and negative shifts are underpinned by somewhat different profiles of values (i.e. increases in trust associated more with detail and credibility; decreases were associated more with a lack of transparency and accuracy). Importantly, trust was not just about *what* was said, but *who* said it, and *how clearly* the reasoning was communicated.

Activity 2: Identifying and Prioritising the Values That Shape Trust in AI Systems

This activity aimed to identify the high-level values that participants believe are most important for establishing trust in hybrid AI tools like the HACID system. Whereas Activity 1 focused on how participants evaluate individual system outputs, Activity 2 asked participants to take a broader, more conceptual view—reflecting on what principles and priorities should guide trustworthy AI design in the climate services context.

The activity consisted of two parts:

1. A **facilitated group discussion** where participants answered the question: *"What would make you trust a tool like the HACID system?"* Participants were encouraged to propose up to five value-based reasons they considered critical for trust. Thirteen unique statements were generated through this discussion.

2. **A pairwise ranking task** conducted using the *AllOurIdeas* platform[38]. Participants compared randomly presented pairs of value statements, selecting the one they believed would more strongly contribute to trust in HACID. This process included the 13 discussion-generated statements and three seeded values from the AI ethics literature: credibility, saliency, and legitimacy.

A total of 117 value pair comparisons were made, and the platform produced a final ranked list based on the proportion of wins versus losses for each statement.

Results: Key Findings and Patterns

The top five values identified through the ranking task reflect an emphasis on transparency, data quality, and system limitations:

1. Communicating uncertainty – AI should admit when it lacks confidence and avoid providing answers it cannot support.
2. Transparency – The system should clearly communicate its own limitations.
3. High-quality data – Inputs should be reliable, well-curated, and scientifically robust.
4. Bias awareness – AI should flag inherent biases in its data or logic.
5. Credibility – Information should come from trusted, authoritative sources.

These results align closely with the themes from Activity 1, reinforcing the importance of transparency, credibility, and solution relevance in shaping user trust—yet also surfaced new values relating to how the system recognises and communicates its own limitations. (i.e. communicating uncertainty and bias awareness).

Conclusions

This workshop explored how trust in AI-supported climate services is shaped not only by the characteristics of system outputs, but also by the broader values that users expect such

[38] https://allourideas.org/

tools to reflect. Rather than assuming a fixed set of abstract values, the activities surfaced domain-specific priorities through two complementary methods: one focused on how trust is built or eroded through direct evaluation of information; the other on eliciting high-level design principles through structured discussion and comparison.

Together, these perspectives offer more than a checklist—they show that trust is not solely dependent on system performance, but emerges through users' engagement with, and interpretation of, the information provided. While participants consistently valued traits like credibility, relevance, and transparency, the way values were expressed varied depending on the framing of the task. This insight has implications both for future participatory methods and for how values are translated into system design.

Converging Themes and Key Values for HACID

Taken together, the results highlight a coherent set of high-level values that should inform the development of HACID and similar AI tools. Across both activities, participants consistently prioritised:

- **Transparency** – How clearly the system shows where its answers come from
- **Data Quality** – Whether the underlying data sources seem reliable
- **Credibility** – Whether the sources of information appear trustworthy
- **Bias Awareness** – Whether the system accounts for different perspectives or potential bias
- **Saliency** – Whether the results feel relevant and useful for real-world decisions
- **Accuracy** – Whether the information provided is factually correct
- **Detail** – Whether the output is sufficiently detailed to support expert decision-making

While communication of uncertainty and community endorsement emerged as important values in Activities 2 and 1 respectively, they were not included in the final list above as they relate more directly to specific system features. Nonetheless, both remain important design considerations moving forward.

These findings reinforce that trust in AI is not just about what the system does, but how it explains itself—its transparency, its ability to communicate limits, and its alignment with user expectations and professional standards. Designing with these values in mind is essential to building meaningful, responsible, and usable AI in the context of climate services.

Evaluating the Participatory Process

Participants felt that their values, ideas, and perspectives were meaningfully considered in the development of the HACID system (see Figure 14). There was a strong sense that the goals of the session were clear, and that the activities provided genuine opportunities to contribute to shaping the system. Participants reported that their contributions were respected and recognised, suggesting that the process felt reciprocal and worthwhile. Participants indicated a clear understanding of how their input would be used and how they would be informed of outcomes, demonstrating strong follow-through and transparency.

The only survey item rated less favourably was related to the composition of the group. Responses suggested that participants did not view the group composition as fully appropriate or inclusive, indicating a need for greater diversity or better communication about the participant selection process in future sessions.

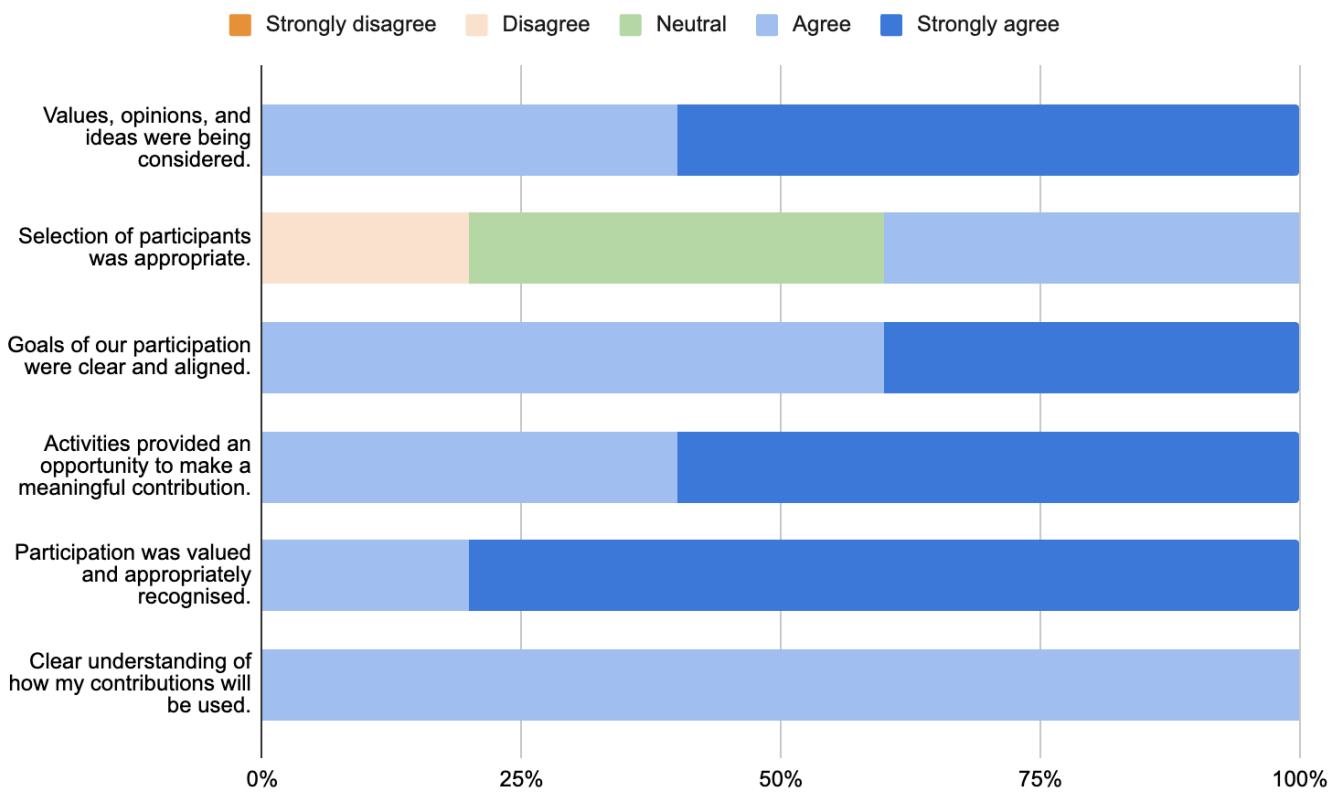


Figure 14: Assessment of the knowledge graph visualisation and values elicitation activities by workshop participants. Likert statements have been shortened for visual clarity. (n=5) Participants were broadly satisfied with the activities as assessed against 6 key participatory dimensions. The one dimension that was least favourably assessed related to the diversity and appropriateness of participants involved.

Five Key Takeaways from the HACID Values Elicitation Workshop

1. **Trust is shaped by both content and context:** Participants evaluated solutions based not only on the quality of information provided, but also on who it came from and how clearly it was explained. Trusted sources, detailed reasoning, and visible relevance consistently led to higher trust ratings.
2. **Transparency and provenance are essential:** Participants repeatedly stressed the importance of being able to trace where information came from and understand how recommendations were generated. This included a desire for source attribution and detailed rationales.

3. **Values are context-sensitive:** Values like credibility and transparency were consistently identified across elicitation methods, while others—such as accuracy or bias awareness—emerged only in specific tasks. Designers should recognise that the values prioritised for AI evaluation may depend on how they are elicited, and consider using multi-method approaches to capture a fuller picture.

4. **Different values drive increases and decreases in trust:** Trust in a solution was most often increased by the detail, credibility, and relevance of the information provided. In contrast, it was most commonly reduced by a lack of transparency, perceived inaccuracy, and insufficient detail—highlighting that different value profiles underpin positive and negative shifts in trust.

5. **Participatory design deepens engagement and surfaces richer insights:** Participants reported that the structured, reflective format helped them think more deeply about AI systems, and increased their willingness to trust and engage with them.

4.2.4 Participatory Evaluation of Use Case 2: Climate Services

Objectives

This participatory evaluation activity aimed to involve the key stakeholders and intended beneficiaries of the HACID-DSS concept in the evaluation of the prototype, and to ensure that it aligns with decision-makers' values, priorities, and concerns, as well as assessing their trust in the system's capabilities and outputs.

Methodology

The participatory evaluation activity consisted of three components:

2. A demo and interactive exercise with the HACID-DSS prototype.
4. A discussion based on the results of a short real-time poll about the values that participants thought the HACID system appears particularly well aligned with; and the values participants thought the HACID system was most poorly aligned with.
5. A post-workshop feedback survey was administered to participants, firstly to assess how well they felt the tool aligned with different values, following the demo and group deliberation. And secondly to evaluate the quality of the activities along five dimensions directly related to good participatory design.

Results

After a clarification of the definition of the term "values" in the context of the workshop, participants were asked to select up to three values that they thought the HACID system demonstrated particularly well based on their interaction with the prototype. The group entered a period of discussion where the following key takeaways emerged:

- **The value of "Transparency"** scored highly in the survey. When discussed, this centred around the need for the tool to clearly show how it has produced the results and recommendations, so that those results can be questioned and understood more

fully. To do this requires a robust level of detail within the recommendations – if the level of detail isn't granular enough then it is not possible to understand how that solution was arrived upon, making it harder to trust the results.

"I'm interested in the fact that people voted for transparency…there's a concern that we might not be able to see how it produces the results."

"Detail matters, because there are many nuances involved when providing a solution."

- **The need for the right contributors to contribute to the solution design from the outset.** In this instance, the users need to feel assured that the climate scientists (and other contributors) were recruited using robust and relevant criteria, and in ways that negate as much as possible the presence of any bias that might skew the recommendations.

"It's important to ensure the right diversity of experts are involved. Without a good range of contributors, the solution may not be as good."

"Some of the solution options may be more valid than others, depending on who submitted them."

- **The importance of Data quality/Data credibility.** During discussions, it emerged that the participants had almost taken these as a given, and that these were viewed as integral to the success of the concept.

"Data credibility and quality are important to me. I can see that we don't currently have the right sources, and securing those is essential."

In the post-workshop survey, participants ranked the HACID-DSS with respect to all seven values that had been discussed during the workshop (Figure 15). In this assessment, the prototype was most highly ranked in the dimensions of Transparency, Bias Awareness, Data Quality and Saliency. At least 5/7 participants agreed that the tool demonstrated these four values. Accuracy, Credibility and Detail were the lowest ranking values, with only 2 or 3 out of 7 participants agreeing that the current prototype aligned with these values.

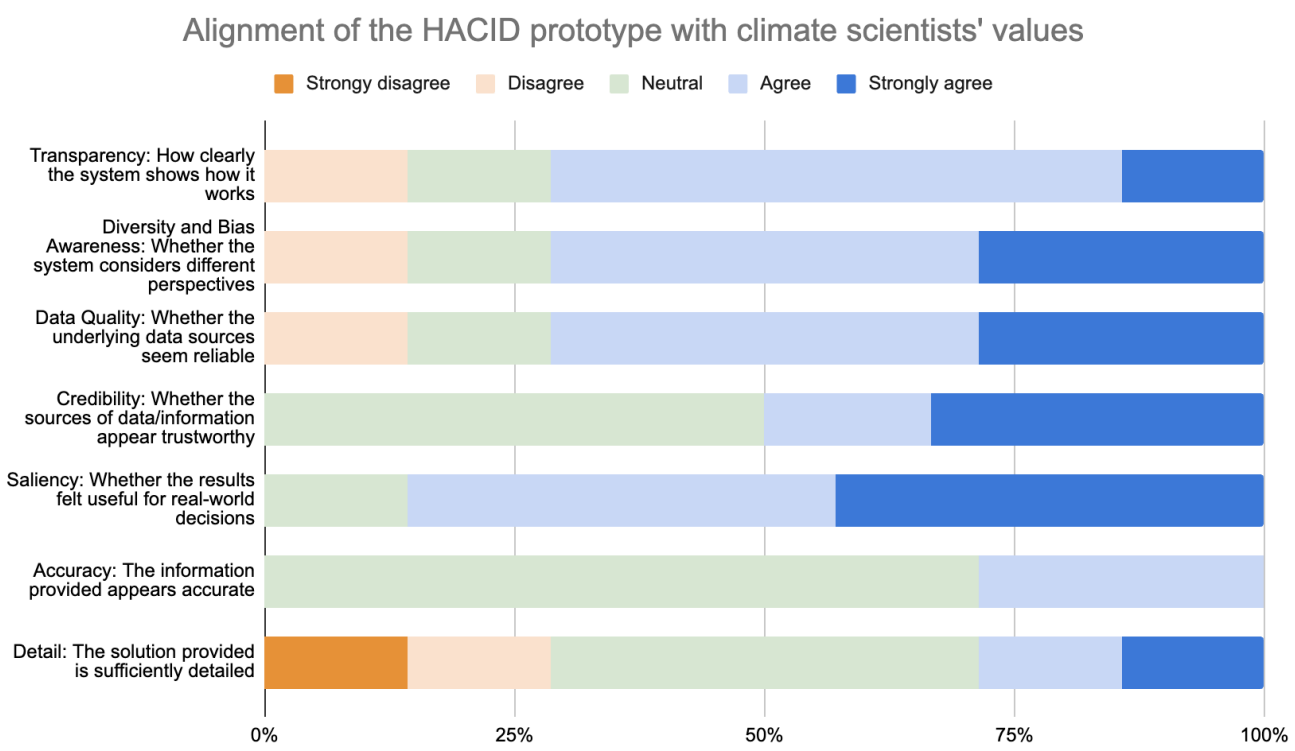


Figure 15: Participants' assessment of how well the current HACID prototype aligns with the professional values prioritised by climate scientists. (n=7)

Conclusion

The climate scientists broadly agreed with the prioritised list of values that had been specified by their peers through the previous participatory activity. During the small-group facilitated deliberation, participants highlighted that the current HACID-DSS was particularly well aligned with Transparency and Bias Awareness. They also emphasised the importance of Data Quality. In the post-workshop survey, these 3 values were also rated highly, with 1 of 7 participants disagreeing that they were well represented in the HACID tool. The survey also showed that participants felt the tool was strongly aligned with Salience, or usefulness to real world decisions. The values that were least well aligned were Detail and Credibility - these are important to think about in the next development phases of the prototype to ensure that the tool gets buy-in from professionals.

Evaluating the Participatory Process

At the end of the workshop, participants were asked to complete an online feedback survey via a google form. The final part of the feedback survey aimed to assess the quality of the activities along five dimensions of good participatory design (see Figure 16). 3 of 5 dimensions met our KPI target. The two dimensions of the participatory experience that were rated less highly were "I was able to express what I think is important for trustworthy AI" (3/7 = Neutral, 4/7 = Agree/Strongly Agree); "I understand how my input will be used" (1/7 = Disagree/Strongly Disagree, 2/7 = Neutral, 4/7 = Agree/Strongly Agree).

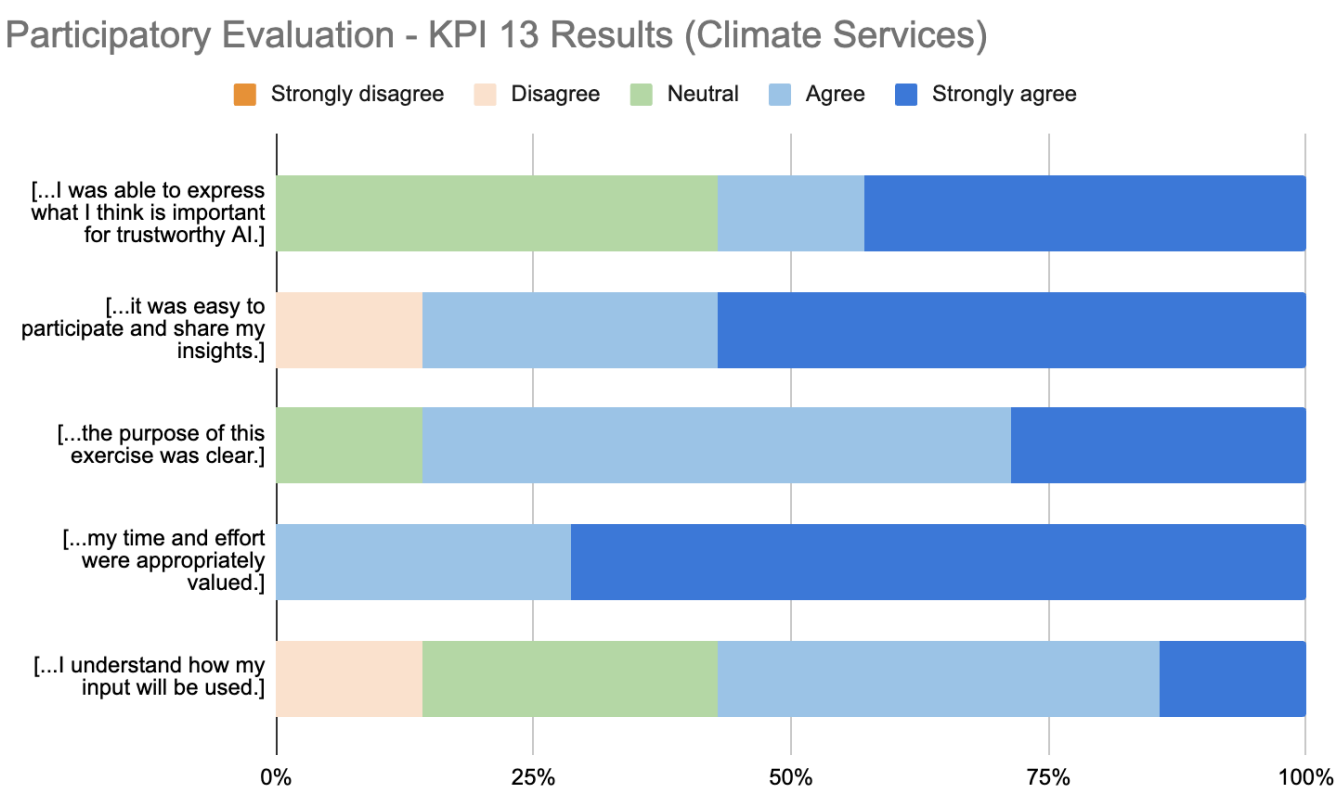


Figure 16: Final assessment of the participatory evaluation (climate services) by workshop participants. 3 of 5 dimensions met our KPI target. (n=7) The two dimensions of the participatory experience that were rated less highly were "I was able to express what I think is important for trustworthy AI" (3/7 = Neutral, 4/7 = Agree/Strongly Agree); "I understand how my input will be used" (1/7 = Disagree/Strongly Disagree, 2/7 = Neutral, 4/7 = Agree/Strongly Agree)

4. Summary assessment against our KPIs

KPI number and description	Progress measure	Target	Results
KPI-10 Process evaluation: How well do the evaluation methods capture the criteria that matter to decision makers?	Validation of the defined evaluation metrics by domain experts	Positive or very positive assessment by ≥ 80% of interviewed experts	In participatory evaluation workshops for both use cases, we started by confirming that the "values" that the prototype was being evaluated against (that had been elicited from others during earlier participatory interventions) actually mattered to the professionals we engaged. In both cases, the prioritised values were validated by all participants.

KPI-11 Output evaluation: To what extent are the HACID-DSS outputs aligned with stakeholder values?	Assessment of the defined evaluation metrics by domain experts	Positive or very positive assessment by ≥ 80% of interviewed experts	**Use case 1** During the participatory evaluation, participants felt that the prototype aligned best with the values of 1) Reducing human bias and 2) Efficiency and least well with Accountability/liability for errors. Overall, participants recognised that the prototype had been designed to align with several of the values that mattered to them but also saw room for improvement in future design iterations.
			Use case 2 During the participatory evaluation, participants felt that the prototype aligned best with the values of Saliency, Transparency, Bias Awareness and Data Quality and least well with Detail, Accuracy and Credibility. Overall, participants recognised that the prototype had been designed to align with several of the values that mattered to them but also saw room for improvement in future design iterations.
KPI-12 Participatory design: Have we introduced new participatory approaches to AI development?	Number of participatory interventions throughout the tool development pipeline	≥ 5 participatory interventions	The total number of participatory activities (led by Nesta) across the HACID project, using a combination of proactive and reactive approaches was 11, with 5 for use case 1 and 6 for use case 2. This does not include smaller-scale, ongoing participatory design and validation activities related to the knowledge graph development led by ISTC-CNR/Met Office or the experiments with clinicians led by Max Planck Institute and Human Dx.
KPI-13 Participatory evaluation: How well did we achieve the goals of participation?	Level of satisfaction reported by participants and relevant stakeholders.	High satisfaction and/or meaningful engagement reported by ≥ 80% of stakeholders.	**Use case 1** <u>Risk Assessment Workshop:</u> More than 80% of participants felt that their values, ideas, and perspectives were meaningfully considered in the development of the HACID system across most dimensions of participation. The lowest ranked criteria related to being able to express their preferences for trustworthy AI. <u>Participatory Evaluation:</u> More than 80% of participants agreed that the activity satisfied all

			of the dimensions of participatory AI good practice. This suggests improvements to the process of participatory engagement as it was the final participatory activity of the project.
			Use case 2 Knowledge Graph Visualisation & Values Elicitation Workshop: Participants felt that their values, ideas, and perspectives were meaningfully considered in the development of the HACID system across most dimensions of participation. The only criteria rated less favourably was about the diversity of participants. This could be due to the relatively low participant number (n=5) and that only 3 different organisations were represented. Participatory Evaluation: More than 80% of participants awarded either an Agree or Neutral across all criteria of participation. The lowest ranked criteria related to being able to express their preferences for trustworthy AI and knowing how their inputs would be used.

Table 5: Summary assessment of participatory approach against our KPIs

5. Key Takeaways and Guidelines for Future Design

1. **Project Integration Challenges**

Coordination Between Technical Development and Participatory Research

Implementing a Participatory AI (PAI) approach in this project surfaced several meaningful tensions and practical challenges. One of the most persistent was coordinating between the technical development and adoption of the HACID tool by commercial partners (HDx) within their platform and the participatory research activities. Several specific constraints proved difficult to navigate. For example, the platform had limited capacity to incorporate survey items, and it was challenging to align participatory research goals and the company's internal priorities or policies. Integrating research components into the platform's development cycle was time-consuming and frequently misaligned with the proposed timeline for research activities. Additionally, collecting critical data (such as demographic information) was difficult due to platform limitations or sensitivity concerns. In contrast, for

the climate services use case where the HACID platform was being developed in parallel with participatory activities, there was much more scope for active collaboration with the technical partners. These tensions underscored the challenge of embedding participatory methods within live, commercially driven development environments.

Balancing Scale and Depth in Participation

Our research employed a mix of different approaches including a large-scale survey and in-depth participatory formats, each with distinct strengths and weaknesses. While large-scale methods efficiently gathered a broad range of public perspectives, they often lacked depth, deliberation opportunities, and the ability to fully inform participants. These approaches, exemplified by our public risk perception survey, prioritise breadth and speed but result in shallower individual responses and limited participant engagement.

In contrast, our more in-depth participatory formats, such as workshops, offered tailored, deliberative experiences that fostered richer reflection, consensus building, and comprehensive participant information. However, these dialogic and participatory engagements yielded less generalisable findings due to smaller sample sizes and their qualitative nature. While our process attempted to bridge this gap by bringing clinicians into dialogue with survey results, this integration was largely reactive rather than a full co-design or joint sensemaking effort. Future research should explore ways to bridge between these methods and scale deliberative approaches while retaining depth (for example through deliberative polling).

2. Methodological Considerations

Developing Robust Values Elicitation Instruments

Values elicitation should make use of criteria elicited using both bottom-up and top down participatory approaches. Findings from our values elicitation workshop demonstrated that, although there were a consistent set of values elicited by our participants across activities, there were some notable/important variations between tasks that focused on evaluation of simulated system output (i.e. bottom-up), and broader discussions of values in AI tools (i.e. top-down). System designers and developers should ensure a diverse set of values elicitation procedures are used to develop instruments/metrics of values alignment.

Distinguishing between Attitudes to AI in General and Tool-Specific Attitudes

As AI becomes increasingly pervasive, participants often bring existing attitudes towards AI to participatory activities. It is crucial to design activities that allow space for airing these general attitudes while also supporting participants to focus on the specific tool's technical attributes and applications. Failing to differentiate can lead to participants struggling to distinguish between their feelings about AI in general and hybrid intelligence tools like HACID.

We recommend using structured, facilitated workshops with scenarios that demonstrate use to bring the application to life. The facilitator should be prepared to answer clarifying

questions about the technical details of the tool. Framing the tool within a hybrid intelligence context also helps participants reflect on the limitations of human advice and crowdsourcing.

3. Engagement & Resource Barriers

Reaching Beyond the 'Usual Suspects'

We also encountered challenges in reaching beyond the 'usual suspects' — individuals already familiar with or predisposed toward participatory processes. Achieving broader and more inclusive outreach requires not only thoughtful recruitment strategies but also sufficient time, dedicated funding, and strong partnerships with trusted organisations. To some extent, this challenge is systemic and reflects the nature of the communities we engaged with. Both climate scientists and clinicians typically come from highly specialised professional domains and often represent more privileged socioeconomic backgrounds (e.g., higher levels of education). Consequently, these groups tend to over-represent certain demographic characteristics, limiting the diversity of perspectives and potentially reinforcing existing inequalities within participatory processes. This underscores the importance of embedding diversity targets within recruitment strategies and explicitly capturing socioeconomic and demographic characteristics of participants to enable process evaluation and continuous improvement.

Financial and Resource Constraints

Another limitation of the Participatory AI approach is its financial cost. High-quality participatory work requires substantial investment — including fair compensation for participants, the design of inclusive materials, and support from experienced facilitators. While we were able to remunerate participants for some activities (for example, medical practitioners), this was not consistently applied across all engagements due to institutional guidelines for certain types of stakeholders.[39] Future research should prioritise consistently funding these essential capacities to ensure the quality and effectiveness of participatory processes.

Engaging Busy Professionals

Finally, engaging busy professionals, such as clinicians or climate scientists, proved particularly challenging. Their availability for in-depth co-design is limited. This meant that several participatory activities sometimes needed to be combined with other activities rather than standalone engagements in order to maximise the time being granted by professionals. This in turn, limited the scope of some of the engagements. What made participation possible in our case were enabling factors like strong partner organisations (e.g., the Met Office), flexible engagement formats, and appropriate financial incentives for recruitment through professional recruitment agencies. Without these, the barrier to meaningful professional involvement would have been significantly higher.

[39] Internal guidelines from partner institutions prevented us from issuing financial incentives to climate scientists for example.

Appendix

Survey items for evaluating KP13

We used the following survey items used to evaluate participation quality, as a measure for KPI13. The survey was developed to align with the recommendations for good practice in participatory AI as outlined in Nesta's Participatory AI framework and Delgado et al's Participatory turn in AI Design. We asked participants to complete the survey for all participatory interventions that followed the initial user research phase. User research activities were not evaluated in the same way because the project KPIs were finalised after the completion of the user research phase.

- *Who defines the process and what counts as success?*
  "...my values, opinions, and ideas were being considered in the design and development of the AI system."

- *What is the intent behind participation?*
  "…the goals of our participation were clear and aligned with improving the AI system." & "…the workshop activities provided an opportunity for me to make a meaningful contribution to the development of the AI system."

- *How will participants be rewarded?*
  "…my participation was valued and appropriately recognised."

  What is the process for closing the project?
  "…I have a clear understanding of how my contributions will be used and how I will receive feedback about the project's outcomes."

- *Whose participation is required?*
  "…the selection of participants was appropriate for the purpose of developing an AI decision support system for climate services."

Summary of clinicians' reflections about the Human Dx app

We provide a summary of key takeaways from two questions posed during the pre-task of the participatory evaluation of Use Case 1.

Scenario: You encounter a complex patient symptom or a unique case where you'd like to get input from a broader network of experienced clinicians.

Action: Navigate to Human Dx's "Community" section (the central button in the menu panel at the bottom of the screen). Explore how you would post a question, browse existing discussions, or contribute to a conversation.

Feedback: Please summarise your impressions here in a few short bullet points / sentences. Note how easy or difficult it is to find relevant insights. What questions do you have?

- **Ease of Use:** Most users found the app easy to navigate for asking questions and exploring case discussions.
- **Search and Filtering Difficulties:** Several users experienced difficulty filtering content by their specialty, finding relevant content, or searching the community for specific topics.
- **AI and Community Feedback:** Users raised questions about the purpose and quantity of AI responses, the convergence score, and the credentials/reliability of human respondents. Some noted a lack of clinical questions in community posts.
- **Specialty Relevance & Features:** The app appeared to be largely medical rather than dental, leading to questions about its value for dental professionals and the possibility of uploading radiographs/photos. Users also inquired about region-based filtering due to differing guidelines.
- **Interface and Functionality:** Some users noted a glitchy/slow interface and a lack of clarity regarding the order of differentials or the ordering of AI responses. One user pointed out an incorrect button description in the instructions.

Scenario: Consider your typical day and how you usually make decisions about cases. Now imagine you're able to integrate the Human Dx app into your process.

Action: Think about how you would switch between your current workflows and the Human Dx app.

Feedback: Please summarise your thoughts in a few short bullet points / sentences. Can you picture yourself using this app? In what scenarios would it be useful?

- **Usefulness for complex or uncertain cases:** The app would be beneficial for cases where clinicians are unsure, need to confirm differential diagnoses, or seek second opinions, especially in critical care or when dealing with unfamiliar presentations.
- **Educational and training purposes:** The app could be valuable for training students and trainees, generating questions, and facilitating department-wide training.
- **Concerns about workflow integration and response time:** Some users expressed concerns about integrating the app into busy daily routines, the speed of responses for urgent cases, and the lack of specialists for their specific fields.
- **Current alternative tools and support:** Several clinicians already use other validated online tools or have access to expert colleagues and multi-disciplinary team (MDT) discussions for complex cases.
- **Suggestions for improvement:** Users suggested a separate section for radiology and diagnostics, a desktop version for easier integration into clinical systems, and a clearer distinction from AI-based tools.

Risk Assessment Survey

Sampling

The table below summarises the total number of responses collected per scenario, the number of those that were valid risk statements–i.e. were not manually coded as unconcerned or unsure, and the number ultimately included in the topic modelling after removing responses deemed outliers by the topic modelling algorithm.

Scenario	Participants (viewed)	Participants (with valid risk statement)	Total Responses	Risk Statements	Used in Topic Modelling	Dropped as Outliers
Rare	152	89	311	152	143	9
Mental	164	124	362	216	204	12
Hormonal	163	119	356	216	196	20
Total	241	–	1029	584	543	41

Topic Modelling of Public Risk Concerns

A topic modelling pipeline was used to identify and categorise free-text risk concerns submitted by members of the public in the risk perception survey. These themes were used to inform risk-ranking exercises in the subsequent clinician workshop.

To generate the stimuli of a ranked list of the general public's perceived risks, a topic modelling pipeline was applied to a subset of early responses (n=241[40]).

Thematic clusters were identified using unsupervised modelling techniques and then refined as categories of risk (e.g. ‘Overreliance on AI for clinical judgement’) by providing the 10 most representative documents (participant identified concerns) to a large language model (chatGPT 4o), and asking it to generate intuitive labels for each cluster based on the statements. This process yielded six distinct risk categories, which were subsequently presented to clinicians during the workshop for reflection and ranking.

Dataset

A total of 241 completed survey responses were available at the internal analysis deadline. Each respondent was presented with two of three clinical scenarios and invited to provide up to three free-text concerns per scenario. After filtering out “no concern”, “unsure”, and ambiguous responses, the final dataset contained 545 discreet risk statements suitable for topic modelling.

Data Preparation

[40] While 286 public responses were ultimately collected, analysis for the workshop was based on an initial subset of 241 submissions, due to internal deadlines and resourcing constraints. All responses were stored, and the additional data may be included in a later analysis.

Free-text risk responses were cleaned and normalised. Responses indicating “no concern” or expressing uncertainty (e.g. “not sure”, “don’t know”) were removed. Remaining entries were combined with accompanying “explanation” text, where available, to provide fuller context. This yielded a corpus of 545 risk statements.

Modelling Approach

To identify underlying themes in the risk data, we applied the BERTopic modelling framework, which combines semantic embeddings with clustering and topic extraction:

- **Embedding**: Texts were embedded using `all-MiniLM-L6-v2`, a fast and effective sentence-transformer model.
- **Dimensionality Reduction**: Embeddings were reduced via UMAP to preserve semantic structure while improving clustering performance.
- **Clustering**: BERTopic used a density-based algorithm to identify natural groupings in the data.

Label Refinement

To improve interpretability and domain relevance, topic labels were manually reviewed and refined using a human-in-the-loop process:

1. The top representative responses from each topic cluster (n=10) were extracted and reviewed.
2. A large language model (chatGPT 4o) was prompted to suggest clear, scenario-independent risk-framed labels.
3. Labels were collaboratively edited to ensure they were: Distinct and non-overlapping, Aligned with qualitative content, Framed in risk language for coherence with the workshop task

This process resulted in six final risk categories:

Topic ID	Final Label
0	Inaccurate or Oversimplified Diagnosis
1	Overreliance on AI for Clinical Judgment
2	Lack of Transparency in Diagnostic Reasoning
3	Misleading or Incomplete Information
4	Breach of Patient Data Privacy
5	Bias and Inequity in AI Decision-Making

Topic -1 (outliers) was excluded from downstream analysis.

Vignettes used for Participatory Risk Assessment

Introduction - HACID tool (video transcript)

Making a medical diagnosis isn't always straightforward. Doctors often face the pressure of making decisions about the best diagnosis or treatment plan — even when time is short, or the case is anything but simple.

To help with this, healthcare institutions are increasingly exploring the use of tools powered by artificial intelligence — or AI. These tools can support diagnosis by spotting patterns in patient data and suggesting possible outcomes based on medical records, clinical guidelines, and similar cases.

But despite the promise, there are still big questions: Is the AI tool fair? Is it safe? And perhaps most importantly... Can doctors — and patients — understand how it reaches its decisions?

That's why researchers are developing new types of AI tools that combine artificial and human intelligence — sometimes called hybrid intelligence. One example is a tool called HACID.

Here's how it works: The HACID tool is designed to work with doctors, not replace them. A clinician can submit a challenging case and question to an online community of medical experts, who review it and suggest possible answers.

The HACID tool also uses AI to get additional advice — based on the patterns it finds in new research findings and historical patient data — to suggest a solution.

At the end, the HACID tool shows a ranked list of answers. The clinician can also explore both the AI solution and the expert answers in detail so they can make a decision about the case.

By combining human expertise with the speed and scale of AI, hybrid tools like HACID offer a powerful way forward. HACID is still being developed — and your views can help shape how it works. By sharing your thoughts, you'll help ensure this tool is built in a way that supports both patients and healthcare professionals.

Scenario 1 - Rare Disease Diagnosis (audio transcript and image)

Elizabeth had been experiencing joint pain, easy bruising, and ongoing fatigue. After several consultations, her symptoms remained unexplained, and her physician, Dr Stevens, began to consider less common possibilities.

To support the diagnostic process, Dr Stevens used HACID, a hybrid AI tool, designed to assist with complex cases. HACID analysed Elizabeth's symptoms by drawing on other experts' input and rare case histories, helping to suggest conditions that might otherwise be overlooked.

One possibility it flagged was Ehlers-Danlos Syndrome (EDS) — a rare connective tissue disorder that often goes undiagnosed or is mistaken for something more common. It's thought to affect between 1 in 5,000 and 1 in 20,000 people and usually requires input from specialist clinics.

Cases like Elizabeth's are especially difficult to diagnose. There may be few experts, limited case data, and a natural focus on more common conditions.

HACID helped Dr Stevens spot the pattern in Elizabeth's symptoms and contributed to the next steps in the diagnostic process. Dr Stevens reviewed the results from the HACID tool carefully and made the final Decision.

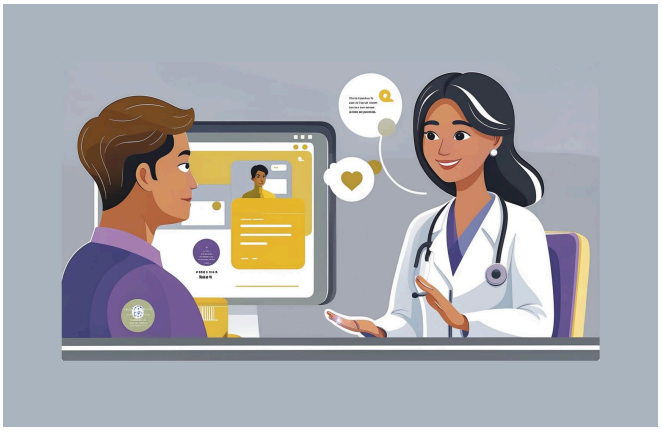
Scenario 2 - Mental Health Diagnosis (audio transcript and image)

Robert had been experiencing anxiety and a persistent low mood. He spoke with his psychiatrist, Dr Parker, who considered whether there might be an underlying condition contributing to his symptoms.To support the diagnostic process, Dr Parker used HACID, a hybrid AI tool designed to assist clinicians.

Using Robert's reported symptoms, HACID compared similar cases using AI and combined advice from several experts to suggest a possible diagnosis.

Mental health diagnoses can be challenging — different clinicians might reach different conclusions, and not everyone has access to a full team of specialists. In this case, HACID gave Dr Parker a second opinion, helping to check whether anything had been missed.

Because of the sensitivity of Robert's situation, it was important to take into account his personal history and current life circumstances. Dr Parker reviewed the result from the HACID tool but made the final diagnosis herself.



Scenario 3 - Hormonal Imbalance (audio transcript and image)

Anna had been experiencing persistent fatigue and mood swings for several months. These symptoms were often viewed through the lens of stress.

Looking for answers, Dr Lawson turned to the HACID tool to help make sense of Anna's symptoms. Based on the patient information she had entered — including Anna's age, ethnicity, and medical history — the system compared her profile with similar cases using AI and also gathered input from different clinical experts. It suggested a list of possible causes. The top two options were a thyroid disorder or early perimenopause.

Hormonal conditions often have overlapping symptoms, and timing can be important — especially when symptoms are linked to hormonal cycles. HACID helped Dr Lawson spot patterns in Anna's symptoms, based on the individual profile provided by the patient, making the diagnosis and treatment relevant for this particular person.

Dr Lawson reviewed the results of the HACID tool, using it as one source of input alongside her own clinical judgement.