HACID - Deliverable

Evaluation workflow and KPIs

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:	D5.1
Due date:	28.02.2025
Nature[1]:	R
Dissemination Level[2]:	PU
Work Package:	WP5
Lead Beneficiary:	NESTA
Contributing Beneficiaries:	CNR, MPG, Met Office, Human Dx

[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
V1	10/02/2025	First draft with input from consortium partners	Christopher Edgar	Nesta, CNR, MPG, MetO, HDX
V2	13/02/2025	Internal review	Aleks Berditchevskaia	Nesta
V3	21/02/2025	Second draft	Christopher Edgar, Ewa Dominiak	Nesta
V4	26/02/2025	Reviewed by CNR (coordinator), consortium partners reviewed specific KPI sections. Feedback and suggested edits provided.	Christopher Edgar, Vito Trianni, Stefan Herzog.	Nesta, CNR, MPG.
V5	27/02/2025	Feedback addressed and final revision	Christopher Edgar	Nesta

Table of contents

1. Introduction

1.1. Evaluation of Hybrid Collective Intelligence Systems

In an era of rapidly expanding information and growing complexity, decision-making in high-stakes domains such as medicine and climate change adaptation management stand to benefit from a synergistic integration of human expertise and advanced technologies.[3] The Hybrid Collective Intelligence for Decision-making project is an initiative dedicated to developing and evaluating Hybrid Collective Intelligence (Hybrid-CI) tools that enhance decision-making by combining human collective intelligence with artificial intelligence (AI).

Research highlights Hybrid-CI's potential to enhance decision-making, increase efficiency, and foster innovation.[4] [5] This is further strengthened by the opportunity to implement participatory frameworks, a choice we are making with HACID to improve the system's alignment with end-user values, enhance system responsiveness to stakeholder priorities, and build trust in AI-assisted decision-making.[6] Evaluating Hybrid-CI systems, such as HACID, must therefore go beyond task-specific benchmarking and technical performance assessments. It also requires socio-technical evaluation—measuring how well these systems align with human values, trust, and usability in real-world contexts.

This is easier said than done, as any evaluation involving AI systems is inherently complex, with significant variation in both the tools used and the performance expectations across and within different domains of application.[7] Moreover, the quick pace of AI development (and particularly general purpose AI/LLMs) means that benchmarks are quickly outdated.[8] [9] This underscores the need for an evaluation plan that is comprehensive enough to capture Hybrid-CI's diverse impacts yet tailored to the specific demands of each domain. Determining what to measure and evaluate requires a nuanced understanding of AI's potential benefits, applications, risks, and challenges within its deployment context. Simply put, AI systems—whether conventional or hybrid—can succeed or fail in multiple ways,

---

[3] Trianni, Vito, et al. "Hybrid Collective Intelligence for Decision Support in Complex Open-Ended Domains." *HHAI 2023: Augmenting Human Intellect*, IOS Press, 2023, pp. 124–37. *ebooks.iospress.nl*, https://doi.org/10.3233/FAIA230079

[4] Madni, Azad M., and Carla C. Madni. "Architectural Framework for Exploring Adaptive Human-Machine Teaming Options in Simulated Dynamic Environments." *Systems*, vol. 6, no. 4, Dec. 2018, p. 44. *www.mdpi.com*, https://doi.org/10.3390/systems6040044.

[5] Dubey, Alpana, et al. "HACO: A Framework for Developing Human-AI Teaming." *Proceedings of the 13th Innovations in Software Engineering Conference (Formerly Known as India Software Engineering Conference)*, Association for Computing Machinery, 2020, pp. 1–9. *ACM Digital Library*, https://doi.org/10.1145/3385032.3385044 .

[6] Berditchevskaia, A., Peach, K., and Malliaraki, E. (2021). Participatory AI for humanitarian innovation: a briefing paper. London: Nesta.

[7] Wizards, Data Science. "How Artificial Intelligence Is Advancing Different Domains?" *Medium*, 7 July 2022, https://medium.com/@datasciencewizards/how-artificial-intelligence-is-advancing-different-domains-92384311334.

[8] "LLM Have Been Set Their Toughest Test yet What Happens When They Break It", https://www.turing.ac.uk/blog/llms-have-been-set-their-toughest-test-yet-what-happens-when-they-beat-it. Accessed 26 Feb. 2025

[9] *AI Index Report 2024 – Artificial Intelligence Index*. https://aiindex.stanford.edu/report/. Accessed 26 Feb. 2025.

depending on what they are designed to do and how they interact with their environment and users.

To illustrate these complexities, we set out the technical and socio-technical evaluation considerations shaping HACID's two primary use-case domains: healthcare and climate services.

1.2. Evaluation of AI systems in Healthcare

At a technical level, evaluating AI tools in healthcare often involves benchmarking how well these systems support tasks related to clinical decision-making. AI is seeing rapid adaptation to assist with a broad range of such tasks, including treatment planning[10], predicting patient disease risk[11], and—of particular relevance to HACID—enhancing diagnostic accuracy, i.e., how precisely an AI system identifies diseases compared to a verified ground-truth diagnosis.[12] Complementing measurement of task-specific performance, research has also explored the alignment between AI and practitioner diagnoses, emphasising the value of ensuring AI recommendations are clinically relevant and effectively integrated with human expertise.[13]

Beyond technical benchmarks focused on clinical decision-making, the evaluation of AI in healthcare must also consider broader system-wide impacts linked to its perceived benefits and risks. For example, AI adoption in health services has the potential to free up clinicians' time, improve workflow efficiency, and enhance patient safety, highlighting the need for evaluation metrics that assess organisational efficiency and risks to patient privacy.[14] [15] In addition, public perception research suggests that while AI integration in healthcare is generally accepted, significant concerns remain around diagnostic accuracy, AI's lack of empathy, and the potential for physicians to become over-reliant on technology.[16] These concerns highlight the importance of evaluation metrics that assess broad metrics such as user trust, emotional response, and AI's impact on clinician skills development.

---

[10] Blasiak, Agata, et al. "CURATE.AI: Optimizing Personalized Medicine with Artificial Intelligence." *SLAS Technology*, vol. 25, no. 2, Apr. 2020, pp. 95–105. *DOI.org (Crossref)*, https://doi.org/10.1177/2472630319890316.

[11] Gameiro, Joana, et al. "Artificial Intelligence in Acute Kidney Injury Risk Prediction." *Journal of Clinical Medicine*, vol. 9, no. 3, Mar. 2020, p. 678. *www.mdpi.com*, https://doi.org/10.3390/jcm9030678.

[12] Ghaffar Nia, Nafiseh, et al. "Evaluation of Artificial Intelligence Techniques in Disease Diagnosis and Prediction." *Discover Artificial Intelligence*, vol. 3, no. 1, 2023, p. 5. *PubMed Central*, https://doi.org/10.1007/s44163-023-00049-5.

[13] Zeltzer, Dan, et al. "Diagnostic Accuracy of Artificial Intelligence in Virtual Primary Care." *Mayo Clinic Proceedings: Digital Health*, vol. 1, no. 4, Dec. 2023, pp. 480–89. *ScienceDirect*, https://doi.org/10.1016/j.mcpdig.2023.08.002.

[14] *The Benefits of the Latest AI Technologies for Patients and Clinicians | HMS Postgraduate Education*. 30 Aug. 2024, https://postgraduateeducation.hms.harvard.edu/trends-medicine/benefits-latest-ai-technologies-patients-clinicians.

[15] Jemma Kwint "AI in Healthcare - 10 Promising Interventions." *NIHR Evidence*, 28 July 2023, https://doi.org/10.3310/nihrevidence_59502.

[16] Young, Albert T., et al. "Patient and General Public Attitudes towards Clinical Artificial Intelligence: A Mixed Methods Systematic Review." *The Lancet Digital Health*, vol. 3, no. 9, Sept. 2021, pp. e599–611. *DOI.org (Crossref)*, https://doi.org/10.1016/S2589-7500(21)00132-1.

Notably, a number of evaluation frameworks for AI in healthcare have been developed, each addressing technical and socio-technical factors to varying degrees. Frameworks such as TRIPOD+AI, DECIDE-AI, SPIRIT-AI, and CONSORT-AI focus primarily on technical validation, methodological rigor, and clinical applicability, ensuring that AI-driven models are robust and reliable.[17] Meanwhile, frameworks like the Translational Evaluation of Healthcare AI (TEHAI) extend beyond technical performance, incorporating capability, utility, and adoption metrics to assess real-world implementation.[18] However, these frameworks do not directly account for Hybrid-CI systems, where AI functions as an active collaborator rather than a standalone decision-making tool. As a result, additional evaluation considerations are required to measure how Hybrid-CI integrates with human expertise, fosters trust, and enhances decision-making in clinical settings.

1.3. Evaluation of AI systems in Climate services

In the domain of climate services, assessing AI performance presents equally complex challenges, requiring a comprehensive evaluation approach that considers both technical and socio-technical factors. The range of potential applications—and evaluation metrics—is vast. AI can contribute to both climate mitigation and adaptation strategies, spanning climate modelling and planning, urban resilience, biodiversity monitoring, energy and transport management, and emissions tracking.[19] Across these domains, technical performance evaluation primarily focuses on the accuracy of climate system predictions,[20] and the anticipated impacts of climate change on economies and infrastructure.[21] [22] [23]

Beyond technical performance, evaluating AI in climate services requires assessing its broader societal impact, including equity, governance, and trust. A key challenge is the digital divide, where AI-driven climate solutions remain concentrated in wealthier nations, leaving low-income and rural communities with limited access to critical climate tools.[24] Additionally,

---

[17] Shiferaw, Kirubel Biruk, et al. “Guidelines and Standard Frameworks for Artificial Intelligence in Medicine: A Systematic Review.” *JAMIA Open*, vol. 8, no. 1, Dec. 2024, p. ooae155. *DOI.org (Crossref)*, https://doi.org/10.1093/jamiaopen/ooae155.

[18] Reddy, Sandeep, et al. “Evaluation Framework to Guide Implementation of AI Systems into Healthcare Settings.” *BMJ Health & Care Informatics*, vol. 28, no. 1, Oct. 2021, p. e100444. *PubMed*, https://doi.org/10.1136/bmjhci-2021-100444.

[19] “Artificial Intelligence for Climate Action in Developing Countries: Opportunities, Challenges and Risks.”, *UNFCCC,,* https://unfccc.int/ttclear/misc_/StaticFiles/gnwoerk_static/AI4climateaction/28da5d97d7824d16b7f68a225c0e3493/a4553e8f70f74be3bc37c929b73d9974.pdf. Accessed 26 Feb. 2025

[20] GraphCast: AI Model for Faster and More Accurate Global Weather Forecasting.” *Google DeepMind*, 25 Feb. 2025, https://deepmind.google/discover/blog/graphcast-ai-model-for-faster-and-more-accurate-global-weather-forecasting/.

[21] Crane-Droesch, Andrew. “Machine Learning Methods for Crop Yield Prediction and Climate Change Impact Assessment in Agriculture.” *Environmental Research Letters*, vol. 13, no. 11, Oct. 2018, p. 114003. *DOI.org (Crossref)*, https://doi.org/10.1088/1748-9326/aae159.

[22] Chakraborty, Debaditya, et al. “Scenario-Based Prediction of Climate Change Impacts on Building Cooling Energy Consumption with Explainable Artificial Intelligence.” *Applied Energy*, vol. 291, June 2021, p. 116807. *ScienceDirect*, https://doi.org/10.1016/j.apenergy.2021.116807

[23] Betz, Trevor, et al. “Machine Learning Model to Predict Impact of Climate Change on Facility Equipment Service Life.” *Building and Environment*, vol. 234, Apr. 2023, p. 110192. *ScienceDirect*, https://doi.org/10.1016/j.buildenv.2023.110192.

[24] “Artificial Intelligence for Climate Action in Developing Countries: Opportunities, Challenges and Risks.”, *UNFCCC,,*

AI models risk bias and misrepresentation, as data gaps—particularly in the Global South—can lead to flawed climate predictions that fail to reflect local realities.[25] Ensuring inclusive AI development requires evaluation metrics that measure regional accessibility, human-AI collaboration, and adaptation to diverse socio-environmental contexts.[26] [27]

While established evaluation frameworks for AI in healthcare provide developed methodologies for assessing technical performance and clinical applicability, frameworks specifically developed for AI in climate services remain relatively immature and are still emerging. Initiatives such as the CLARA project[28] and efforts by the Green Climate Fund[29] [30] have begun addressing this gap by proposing evaluation criteria that incorporate quantitative verification, user perceptions, and real-world implementation metrics. However, a standardised and robust evaluation framework tailored to Hybrid-CI applications in climate services is yet to be developed, underscoring the need for further progress in this area.

Governance and ethical considerations are equally crucial in the development and deployment of AI within climate services. Without clear oversight, AI could inadvertently reinforce existing inequalities, contribute to misinterpretation of climate data, or even be misapplied in ways that do not align with climate resilience goals. Evaluating AI's role in climate services must therefore extend beyond technical precision to assess transparency, accountability, and public trust in AI-driven climate insights.[31]

1.4. Domain-Agnostic Hybrid-CI Evaluation Frameworks

A range of domain-agnostic evaluation frameworks have been proposed to assess hybrid or 'human-in-the-loop' AI system performance. While some frameworks prioritise technical robustness, others focus on human interaction, ethical considerations, and policy alignment.

---

https://unfccc.int/ttclear/misc_/StaticFiles/gnwoerk_static/AI4climateaction/28da5d97d7824d16b7f68a225c0e3493/a4553e8f70f74be3bc37c929b73d9974.pdf. Accessed 26 Feb. 2025

[25] Debnath, Ramit, et al. "Harnessing Human and Machine Intelligence for Planetary-Level Climate Action." *Npj Climate Action*, vol. 2, no. 1, Aug. 2023, pp. 1–11. *www.nature.com*, https://doi.org/10.1038/s44168-023-00056-3.

[26] "Artificial Intelligence for Climate Action in Developing Countries: Opportunities, Challenges and Risks.", UNFCC, https://unfccc.int/ttclear/misc_/StaticFiles/gnwoerk_static/AI4climateaction/28da5d97d7824d16b7f68a225c0e3493/a4553e8f70f74be3bc37c929b73d9974.pdf. Accessed 26 Feb. 2025

[27] Debnath, Ramit, et al. "Harnessing Human and Machine Intelligence for Planetary-Level Climate Action." *Npj Climate Action*, vol. 2, no. 1, Aug. 2023, pp. 1–11. *www.nature.com*, https://doi.org/10.1038/s44168-023-00056-3.

[28] https://www.clara-project.eu/. Accessed 26 Feb. 2025.

[29] "Brief Scoping Study on the Use of Artificial Intelligence in Climate Change Evaluations.", *Green Climate Fund,* https://ieu.greenclimate.fund/sites/default/files/event/ai-scoping-study-brief-final.pdf. Accessed 26 Feb. 2025

[30] Fund, Independent Evaluation Unit |. Green Climate. "Ensuring Transparency and Accountability in AI-Driven Climate Evaluations." *Independent Evaluation Unit | Green Climate Fund*, 31 Jan. 2025, https://ieu.greenclimate.fund/news/ensuring-transparency-and-accountability-ai-driven-climate-evaluations.

[31] Gentine, Pierre, Geneva List, Kyoko Thompson, Theresa Pardo, Xin Li, George Berg, Lauren Bennett, et al., Landscape Assessment of AI for Climate and Nature. (May 2024). Available at bezosearthfund.org/ai-climate-nature.

We will now highlight two of these frameworks—each with a significant focus on hybrid human-AI collaboration—to illustrate the broader landscape of evaluation methodologies relevant to HACID's application domains. This will contextualise the complex challenges of assessing Hybrid-CI systems and highlight key considerations in structuring robust evaluation metrics.

1.4.1. The Human-AI Collaboration Evaluation Framework

The Human-AI Collaboration Evaluation (HAIC) framework[32] is designed to assess collaborative AI systems, where humans and AI interact in shared decision-making processes. Unlike traditional AI evaluation methods that focus primarily on accuracy, efficiency, and automation, HAIC emphasises the dynamic and reciprocal relationship between human and AI partners. It accounts for both quantitative and qualitative aspects of collaboration, ensuring that AI is not only effective but also integrates with human workflows.

The framework introduces a decision-tree model, which can be used to design an evaluation approach—i.e. to select relevant evaluation metrics—depending on the mode of Human-AI interaction in question:

- AI-Centric Mode – AI leads decision-making, with minimal human intervention.
- Human-Centric Mode – AI functions as an assistive tool, enhancing human decision-making.
- Symbiotic Mode – a balanced partnership where AI and humans mutually adapt and contribute to decisions.

Each mode of collaboration is assessed through three primary evaluation factors:

- Goals: Evaluates alignment between AI objectives (e.g., accuracy) and human objectives (e.g., usability and efficiency).
- Interaction: Measures communication effectiveness, trust, and feedback between AI and humans.
- Task Allocation: Examines the division of labor, adaptability, and decision-sharing mechanisms.

By combining technical performance measures (e.g. accuracy, task efficiency) with socio-technical considerations (e.g. ethical concerns, transparency), the HAIC framework provides a significant step towards a more comprehensive evaluation approach that aligns with aspects of HACID's Hybrid-CI decision-making goals. However, one limitation of the HAIC is its narrow treatment of socio-technical factors—broader issues related to user trust, values, and participatory engagement are addressed only as a single subfactor within the 'Interaction' dimension. To ensure a comprehensive evaluation of HACID's participatory component, we must look to other established frameworks that provide deeper insights into socio-technical performance criteria.

[32] Fragiadakis, George, et al. *Evaluating Human-AI Collaboration: A Review and Methodological Framework*. arXiv:2407.19098, arXiv, 9 July 2024. *arXiv.org*, https://doi.org/10.48550/arXiv.2407.19098.

1.4.2. Participatory AI Framework

The Participatory AI for Humanitarian Innovation framework[33], developed by Nesta, provides a structured approach for integrating participatory methods into AI development. Although the framework as developed for evaluation of Hybrid-CI in the humanitarian sector, its core framework is broadly applicable to hybrid-CI evaluation across domains, and is particularly relevant to HACID, as it aligns with the goal of ensuring that Hybrid-CI systems reflect the needs and values of the people they serve.

The framework aims to operationalise participatory AI by incorporating a diverse range of key stakeholders (e.g. affected communities, end-users, policy makers) into the AI development processes. This ensures that AI systems are not only technically effective but also aligned with local knowledge, ethical considerations, and social contexts.

At its core, the framework proposes five key design questions that guide the participatory process:

- Who defines the process and what counts as success? – Determining who has control over decision-making and how success is measured.
- Whose participation is required? – Identifying stakeholders and ensuring inclusive engagement.
- What is the intent behind participation? – Clarifying whether participation is for improving AI accuracy, trust-building, or broader empowerment.
- How will participants be rewarded? – Considering incentives and ensuring that engagement is mutually beneficial.
- What is the process for closing the project? – Ensuring that participatory AI initiatives conclude with transparency and accountability.

The framework highlights different levels of participatory design (consultation, contribution, collaboration, and co-creation), acknowledging that meaningful participation can range from data contribution to full co-ownership of AI systems.

This framework is especially relevant to Hybrid-CI systems such as HACID, as it prioritises collaboration and co-creation, ensuring that AI complements and enhances human decision-making rather than merely automating processes. By addressing trust, governance, and alignment with user priorities, it provides an important reference point for HACID's evaluation framework.

1.5. Foundational Principles Across AI Evaluation Frameworks

While individual frameworks such as HAIC and Nesta's Participatory AI framework provide structured approaches for evaluating Hybrid-CI systems, they exist within a broader landscape of AI evaluation principles. Across the literature, a consistent set of core values is

[33] "Participatory AI for Humanitarian Innovation: A Briefing Paper." *Nesta*, https://www.nesta.org.uk/report/participatory-ai-humanitarian-innovation-briefing-paper/. Accessed 26 Feb. 2025.

emerging as integral for the development of trustworthy AI.[34] [35] [36] However, achieving consensus on how these values interrelate and influence one another remains a challenge.[37] Various efforts have attempted to synthesise these principles into cohesive evaluation frameworks, but a standardised framework or examples of best practice have yet to emerge.

Despite this, significant progress has been made in synthesising AI evaluation principles, offering valuable insights for the development of HACID's evaluation criteria. For example, a recent systematic review[38] mapped evaluation criteria to the EU's Seven Trustworthy AI Principles[39], marking an important step in aligning AI evaluation with real-world policy guidance. These principles include:

- Fairness – AI must operate without discrimination, ensuring equitable treatment and mitigating bias in data and algorithms.
- Transparency – AI systems should be understandable and interpretable, allowing users to see how and why decisions are made.
- Human Agency and Oversight – AI should support, not replace, human decision-making, ensuring clear intervention and accountability.
- Privacy and Data Governance – AI must protect user data, comply with privacy laws, and ensure ethical data collection and security.
- Technical Robustness and Safety – AI should be secure, reliable, and resilient, minimising risks from system failures or attacks.
- Accountability – There must be clear responsibility for AI decisions, with mechanisms for tracing, auditing, and correcting errors.
- Societal and Environmental Well-being – AI should promote sustainability, ethical deployment, and positive societal impact.

HACID's hybrid nature enables real-time participatory evaluation, integrating stakeholder values to assess explainability, transparency, and ethical alignment—helping monitor trust in its capabilities and outputs. In high-stakes decision-making, where HACID will play a critical role, misalignment with human values can be as problematic as technical flaws (e.g. such as systematic bias in medical diagnosis[40]). However, translating abstract principles like accountability and safety into measurable evaluation metrics remains a challenge.

---

[34] Balasubramaniam, Nagadivya, et al. "Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements." *Information and Software Technology*, vol. 159, July 2023, p. 107197. *ScienceDirect*, https://doi.org/10.1016/j.infsof.2023.107197.

[35] Chander, Bhanu, et al. "Toward Trustworthy Artificial Intelligence (TAI) in the Context of Explainability and Robustness." *ACM Comput. Surv.*, vol. 57, no. 6, Feb. 2025, p. 144:1-144:49. *ACM Digital Library*, https://doi.org/10.1145/3675392.

[36] Chamola, Vinay, et al. "A Review of Trustworthy and Explainable Artificial Intelligence (XAI)." *IEEE Access*, vol. 11, 2023, pp. 78994–9015. *IEEE Xplore*, https://doi.org/10.1109/ACCESS.2023.3294569

[37] Knowles, Bran, et al. *The Many Facets of Trust in AI: Formalizing the Relation Between Trust and Fairness, Accountability, and Transparency*. arXiv:2208.00681, arXiv, 1 Aug. 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2208.00681.

[38] McCormack, Louise, and Malika Bendechache. "A Comprehensive Survey and Classification of Evaluation Criteria for Trustworthy Artificial Intelligence." *AI and Ethics*, Oct. 2024. *Springer Link*, https://doi.org/10.1007/s43681-024-00590-8.

[39] *Ethics Guidelines for Trustworthy AI | Shaping Europe's Digital Future*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed 26 Feb. 2025.

[40] Stanley, Emma A. M., et al. "Towards Objective and Systematic Evaluation of Bias in Artificial Intelligence for Medical Imaging." *Journal of the American Medical Informatics Association*, vol. 31, no. 11, Nov. 2024, pp. 2613–21. *DOI.org (Crossref)*, https://doi.org/10.1093/jamia/ocae165.

Key challenges include:

- **Clarifying responsibility** when AI assists or automates decision-making.
- **Ensuring transparency** so AI explanations remain interpretable.
- **Mitigating risks** related to misinformation and unintended biases.

1.6. Towards an Evaluation Framework for HACID

While numerous evaluation frameworks exist across domains, none fully capture the unique challenges of Hybrid-CI evaluation, where AI operates as a collaborator rather than an autonomous system. Existing models—whether focused on technical performance, participatory AI, or ethical principles—offer valuable insights but do not comprehensively account for the interaction between AI and human expertise in complex decision-making environments.

Moreover, the breadth of AI's applications in healthcare and climate services means that a one-size-fits-all approach is insufficient. Hybrid-CI evaluation must go beyond accuracy and efficiency, incorporating metrics relating to trust, governance, and human-AI interaction while also addressing the distinct demands of each use case. To achieve this, the HACID consortium has developed 13 Key Performance Indicators (KPIs), designed to assess Hybrid-CI effectiveness, trustworthiness, and impact in decision-making. These KPIs operationalise the principles and considerations discussed in this section, providing a structured approach for evaluating the technical and socio-technical dimensions of Hybrid-CI while ensuring adaptability to different application domains. The following sections provide an overview of the evaluation framework and how it relates to the components of the HACID system, and then details these 13 KPIs, outlining their rationale, methodological approach, and intended contribution to the robust evaluation of Hybrid-CI systems

2. Overview of the 13 KPIs for the HACID system

Building on the foundational principles and evaluation frameworks discussed in the previous section, this section outlines the 13 Key Performance Indicators developed for the HACID evaluation framework. These KPIs provide a structured approach to assessing HACID's hybrid decision-support system, balancing technical performance with socio-technical considerations, tailored to the requirements of its domains of application.

To ensure consistency and comparability, each KPI is documented using a standardised format that includes:

- **Measurement Approach** – Describing the methodology, data sources, and key metrics used.
- **KPI Details** – Outlining the relevant objective within the HACID project.
- **Limitations** – Identifying constraints, challenges, and areas requiring further refinement.

As a point of reference, Table 2.1 provides a structured, 'at-a-glance' summary of the 13 Key Performance Indicators (KPIs) developed for the HACID evaluation framework. Each KPI is presented with its corresponding measurement focus (progress), target values, relevant project objective (OBJ), and lead organisation(s) responsible for its implementation and assessment.

Table 2.1: Summary of HACID Evaluation Framework's 13 Key Performance Indicators (KPIs)

KPI number and description	Progress measure	Target	Objective	Lead Organisation
KPI-1 Improved diagnostics platform: How many features of the HACID technology have been tested in the experimental version of the HDx platform?	Number of HACID-DSS features integrated in the backend and frontend.	≥ 3 features in the backend ≥ 1 features in the frontend	OBJ1.1	HDx
KPI-2 Expert participation: How diverse is the set of experts involved in the building of the HACID-DSS for climate services?	Number of experts and organisations recruited for providing feedback	≥ 15 experts from ≥ 5 organisations	OBJ1.2	MetO, Nesta
KPI-3 Knowledge engineering: How	a) Classification accuracy of NLP	a) F1-score >= .7 for automatic	OBJ2	CNR

well do the components of the knowledge graph cover the problem space?	models for automatic knowledge extraction; b) Knowledge Graph coverage (system usability scale).	knowledge extraction; b) Accuracy@10 >= .75 for competency questions converted to SPARQL queries and executed against the Knowledge Graph.		
KPI-4 Evidence refinement: how much does the system engage users?	Ability of the HACID-DSS to elicit substantive feedback from users	80% satisfaction with data retrieved.	OBJ3	CNR
KPI-5 Collective accuracy: how often does the HACID-DSS include the correct solution as one of the top-ranked solutions in its proposed solution set?	Ability of the collective to identify the correct solution with high probability and reliability.	40% accuracy increase as compared to individual solutions and relative to the maximum possible improvement (i.e. 40% * [100% - average individual. accuracy]).	OBJ4	MPG
KPI-6 Collective effectiveness: How effective is the HACID-DSS in providing actionable outcomes?	Expert-estimated effectiveness of the collective solution in pointing towards the correct solution.	Perceived usefulness for climate services by >70% practitioners.	OBJ4	MPG MetO, CNR.
KPI-7 Collective efficiency: How efficient is the HACID-DSS in terms of decision costs?	Costs associated with obtaining a collective solution with given target accuracy under best conditions.	Significant reduction of costs with respect to baseline aggregation methods.	OBJ4	MPG
KPI-8 Metadata boosts: How do individual confidence, response times, written justifications, and expertise affect collective solutions?	How accuracy increases from the inclusion of meta information from experts, under best conditions.	50% accuracy increase as compared to individual solutions and relative to the maximum possible improvement (i.e. 50% * [100% - average individual. accuracy]).	OBJ5	MPG, CNR, HDx
KPI-9 Social information	Improved accuracy from suitably	Non-negative impact on accuracy,	OBJ5	MPG, CNR

boosts: How does social information affect collective solutions?	exposing experts to social information	compared to baseline with little to no loss in efficiency		
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	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	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	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	Nesta, CNR HDx, MetO

Note: CNR = Consiglio Nazionale delle Ricerche; HDx = Human Diagnosis Project; MetO = Met Office; MPG = Max Planck Institute

The HACID evaluation framework was designed to capture both technical and socio-technical dimensions of Hybrid-CI performance, ensuring a tailored, comprehensive assessment of system performance across two real-world decision-making contexts (health and climate change adaptation management). The 13 Key Performance Indicators developed within HACID span three broad categories: (1) technical implementation and performance (KPIs 1, 3, 4); (2) decision-support effectiveness (KPIs 5–9); and (3) participatory elements, including stakeholder values and risk considerations (KPIs 2, 10–13). See FIgure 2.1 for an overview of how the KPIs have been grouped. This multi-faceted structure enables HACID to assess system performance in a way that extends beyond technical AI-performance benchmarks, integrating consideration of decision-support effectiveness and socio-technical considerations such as human-AI interaction, stakeholder alignment, and participatory governance.

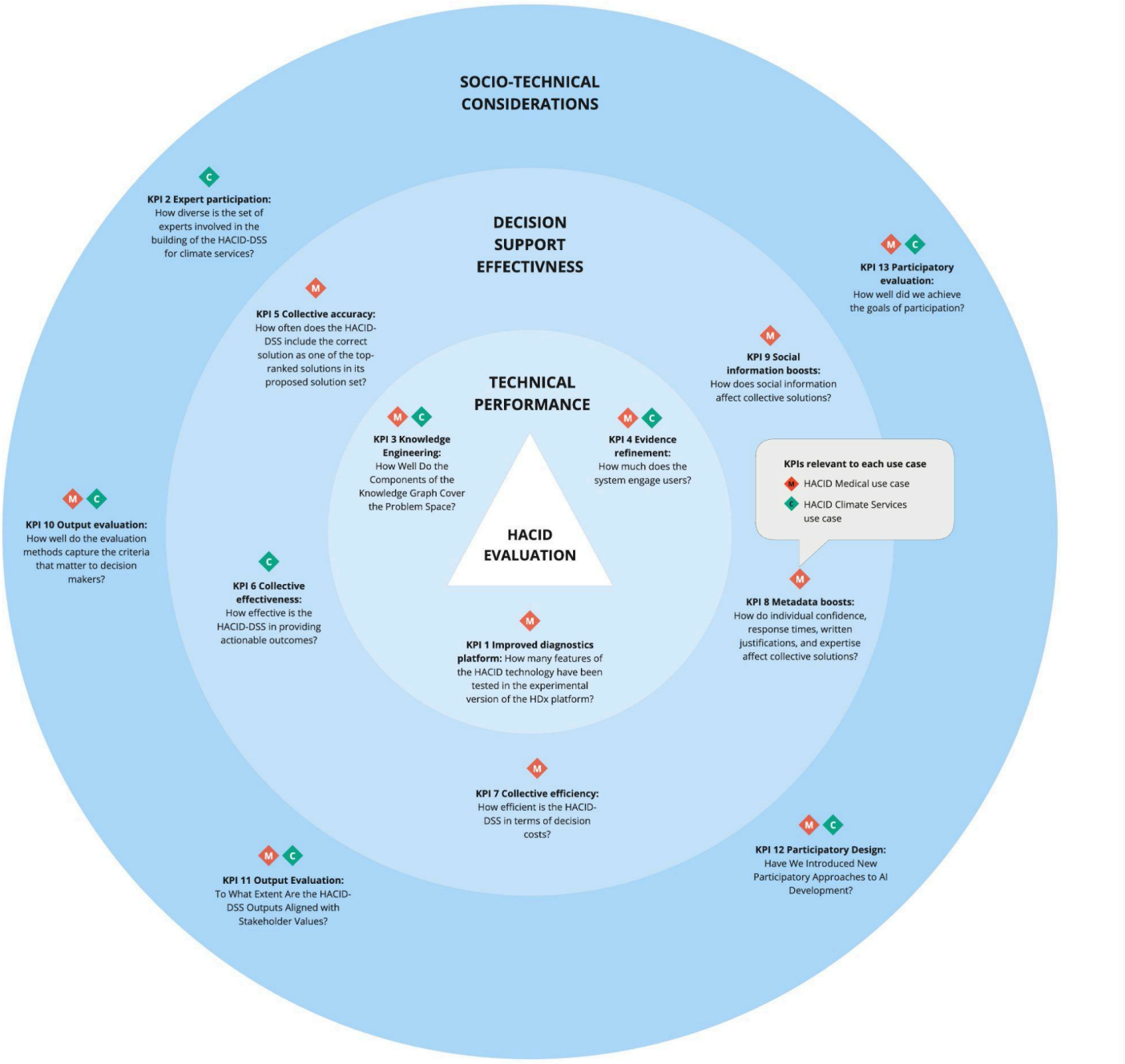

Figure 2.1. Overview of the HACID Evaluation Workflow

2.1. KPI 1 Improved diagnostics platform

How many features of the HACID technology have been tested in the experimental version of the HDx platform?

2.1.1. Relevant Objective

OBJ1.1 Crowd-sourcing medical diagnostics: Exploit hybrid collective intelligence to support physicians in making better diagnoses.

The goal of HACID is to instantiate the DSS in an online platform, adapting the pipeline to the specific needs of the medical diagnostics community, so that crowdsourced diagnoses can be obtained from all over the world.

2.1.2. Measurement Approach

Methodology for Measurement

Status: Complete

Metrics Used

The goal of this KPI is to evaluate the integration of the HACID-DSS technology within the experimental version of the Human Dx platform. Ensuring core functionalities are implemented and tested in both backend and frontend components is essential for system-wide validation.

- **Primary Metric:** Number of HACID-DSS features integrated and tested.
- **Measurement Criteria:** Features tested within both backend and frontend infrastructure.
- **Targets:**
    - **Backend:** ≥ 3 features integrated and tested.
    - **Frontend:** ≥ 1 feature integrated and tested.

Data Sources

The primary data source for this KPI is the Human Dx platform's source code, which provides verifiable records of all integrated HACID-DSS features in both the backend and frontend.

Procedure

The procedure for measuring this KPI involves three key steps. First, clear criteria are established to define which technologies qualify as HACID-DSS features, ensuring consistency in identifying relevant components. Next, a systematic review of the Human Dx platform's source code is conducted to identify and verify the implementation of these features within both the backend and frontend. Finally, each identified feature undergoes

functional testing, with results documented to confirm successful integration and operational performance within the platform.

Rationale for Approach

The number of HACID features integrated and tested within the Human Dx platform serves as an objective measure of both technical progress and real-world applicability. A feature can only be integrated after meeting rigorous criteria: a) well-defined functionality, b) compatibility with the Human Dx platform architecture, and c) alignment with user experience standards. Thus, this KPI also acts as an indicator for broader HACID development, as successful integration requires not only conceptual development but also operational readiness in a clinical decision support context.

2.1.3. Limitations

This KPI **ensures** early identification of implementation challenges; however, defining what qualifies as a “feature” **remains** ambiguous. This **leads** to inconsistencies in measuring progress and **requires** clear criteria to maintain evaluation consistency.

2.2. KPI 2 Expert participation

How diverse is the set of experts involved in the building of the HACID-DSS for climate services?

2.2.1. Relevant Objective

OBJ1.2 Improving climate services: Exploit hybrid collective intelligence to support adaptation management in urban areas.

The goal of HACID is to develop a DSS to help a diverse set of experts provide support to policy makers in managing and building resilient cities able to mitigate the effects of climate change.

2.2.2. Measurement Approach

Methodology for Measurement

Status: Complete

Metrics Used

This KPI measures how effectively experts and institutions are engaged in the system's development.

- **Primary Metric:** Number of experts and organisations engaged in providing feedback.
- **Measurement Criteria:** Participation count from distinct institutions and domains.
- **Target:** ≥ 15 experts from ≥ 5 organisations.

Data Sources

Data is collected through varied engagement methods to capture diverse expertise, including:

- **Structured & unstructured interviews** with domain specialists.
- **Online surveys & email communications** to gather standardised data efficiently.
- **In-person & virtual workshops** to facilitate **cross-disciplinary discussions** and ensure broad participation.

Procedure

The engagement process is designed to ensure meaningful participation across activities. For each method, clear research objectives and questions are established to guide data collection. Emails and surveys are distributed with detailed instructions and follow-up reminders to maximise response rates, while one-to-one interviews are conducted flexibly using a semi-structured guide for consistency.

Iterative interactions are organised throughout the project to support continuous learning and co-development. Workshops are carefully planned, with tailored agendas and pre-workshop materials to encourage active participation. Recruitment strategies focus on ensuring a diverse and representative sample, reflecting the range of expertise required for effective climate services development.

Rationale for Approach

Tracking institutional and domain diversity is crucial, as climate adaptation requires interdisciplinary expertise. Effective climate services rely on a range of perspectives—from climate science and urban planning to policy and community engagement. Ensuring representation across multiple institutions and disciplines supports decision-making that is both scientifically rigorous and contextually relevant.

Our mixed-methods approach, which includes use of surveys, interviews, and workshops (online and in person) accommodates different participation preferences and availability, while also capturing in-depth insights. Diversity was systematically tracked across all activities via surveys, registration forms, and direct outreach, ensuring a representative sample of expert input.

2.2.3. Limitations

This approach effectively captures diverse expertise; however, participant fatigue from repeated engagements affects response rates. Additionally, some activities experience lower-than-expected turnout, likely due to competing commitments.

2.3. KPI 3 Knowledge Engineering

How well do the components of the knowledge graph cover the problem space?

2.3.1. Relevant Objective

OBJ2 Knowledge Engineering: Synthesise available evidence into structured knowledge that can be accessed and processed to support decision-making.

The goal of HACID is to provide a participatory approach to evidence synthesis, co-creating knowledge representation and (semi-)automatically populating the knowledge base to support reasoning and decision-making.

2.3.2. Measurement Approach

Methodology for Measurement

Status (KPI13-a): In progress

Status (KPI13-b): Complete

Metrics Used

HACID-DSS relies on two core capabilities for effective decision support: (1) extracting structured knowledge from unstructured text and (2) ensuring the knowledge graph accurately represents relevant concepts and relationships. This KPI is divided into KPI 3a (Knowledge Extraction) and KPI 3b (KG coverage through Competency Question Verification), each addressing a different aspect of knowledge processing.

KPI 3a: Knowledge Extraction

The ability to extract structured information from unstructured text is critical for HACID-DSS. This involves identifying key entities, relationships, and semantic structures to build a usable knowledge base. To evaluate this process, the system's outputs are compared against gold-standard annotations, focusing on both accuracy (precision) and completeness (recall). Additionally, the number of correctly identified semantic triples (subject-predicate-object relationships) is assessed to ensure high-quality knowledge representation.

- **Primary Metric:** Accuracy of extracted knowledge.
- **Measurement Criteria (and target):**
    - Precision (**target:** ≥ 0.75): Measures how accurately relevant information is extracted while minimising false positives.
    - Recall (**target:** ≥ 0.60): Evaluates the proportion of relevant information successfully retrieved.
    - F1 Score (**target:** ≥ 0.70): Provides a balanced measure between precision and recall.

- Hit Triples Ratio (**target:** ≥ 0.65): Measures the proportion of correctly identified semantic relationships to total extracted triples.

KPI 3b: Knowledge Graph Coverage Through Competency Question Verification

A well-structured knowledge graph should not only store extracted information but also support complex reasoning tasks. To evaluate how well HACID-DSS enables knowledge retrieval and structured inference, Competency Question verification is used. This method assesses whether the system can correctly retrieve relevant information when queries are posed in SPARQL, the standard language for querying knowledge graphs. The Accuracy@10 metric is used to measure whether the expected answer appears within the top 10 results, reflecting how well the system structures and connects knowledge.

- **Primary Metric:** Competency Question verification performance.
- **Measurement Criteria:** Accuracy@10 – proportion of SPARQL queries where the expected answer is within the top 10 results.
- **Target:** Accuracy@10 ≥ 0.75 (i.e., at least 75% of competency question verification test cases must return the expected answer within the first 10 results).

Data Sources

Regarding **KPI3-a**, the evaluation framework utilises two different datasets which provide annotated entities and can be exploited for evaluating parts of the knowledge extraction process.

- BC5CDR[41]: A biomedical benchmark dataset containing 50 abstracts with gold-standard annotations for disease and chemical entities.
- MIMIC-IV[42]: A comprehensive clinical dataset with SNOMED Clinical Terms ontology concept annotations, comprising 204 discharge notes with 51,574 annotations spanning 5,336 distinct concepts.

Concerning **KPI3-b**, the primary data sources used include:

- Competency questions, as collected from domain experts (as a result of focus groups and other use case definition and requirements elicitation processes) and from the analysis of domain data during the ontology design phase;
- Test case specifications, formalised according to the OWUnit vocabulary, where SPARQL query mappings, target KG, expected answers and test environment data are specified
- Test case execution results and logs, as produced by the OWLUnit tool, where the outcome of test execution runs is available

Procedure

KPI3-a

The evaluation of KPI3-a follows a structured methodology.

[41] http://www.biocreative.org/tasks/biocreative-v/track-3-cdr. Accessed 26 Feb. 2025.
[42] https://physionet.org/content/snomed-ct-entity-challenge/1.0.0/. Accessed 26 Feb. 2025.

- **Dataset preparation**: Preprocessing of two biomedical datasets, BC5CDR and MIMIC-IV (see above), to ensure compatibility with the (RAG) knowledge extraction pipeline. The BC5CDR dataset is employed for entity-pair extraction, linking diseases and chemicals in text, whereas the MIMIC-IV dataset focuses on medical concept extraction, aligning extracted terms with SNOMED Clinical Terms[43] annotations. These datasets are formatted to ensure consistency in structure and annotation, enabling reliable comparison between extracted data and ground truth labels.
  Knowledge extraction: The system is tested using two RAG-based configurations. The All-in-One RAG model performs end-to-end structured information extraction, while the Extractor+RAG approach adopts a modular strategy by separating retrieval from structured generation. The extraction pipeline processes text to generate structured outputs, including entity-type pairs that classify medical concepts, semantic triplets that capture relationships between entities, and SNOMED-compliant descriptions that map extracted entities to standardised medical concepts. Systematic testing across various retrieval parameters, such as Top-K values, helps determine optimal extraction configurations.
- **Performance assessment:** Involves a dual evaluation strategy. Quantitative evaluation employs precision, recall, and F1 score to measure extraction accuracy using annotated datasets, while the hit triples ratio is used to assess the correctness of extracted semantic relationships. Additionally, qualitative evaluation includes a large language model-based review to check contextual accuracy and SNOMED Clinical Terms validation to ensure alignment with standardised medical terminology. The insights from these evaluations drive iterative refinements, improving the system's robustness and accuracy over time.

KPI3-b

The methodology for KPI3-b focuses on defining, executing, and assessing competency question verification test cases to ensure the Knowledge Graph effectively answers domain-specific queries.

- **Defining competency questions:** Competency questions are formulated to reflect the ontology's knowledge representation requirements. These questions, derived from domain experts and ontology design analysis, act as benchmarks for what the KG should be able to answer. Each competency question corresponds to an expected information retrieval task that the KG is designed to fulfil. To assess the KG's performance, a SPARQL query is formulated for each competency question. These queries are designed to retrieve the relevant information from the KG using the ontology's structure, and each query is paired with an expected result that represents the correct answer. This ensures that the system's responses can be systematically evaluated.
- **Verification:** involves executing each SPARQL query against the target KG and comparing the actual results with predefined expectations. Queries are tested using the OWLUnit framework[44], which structures test cases in RDF. The execution results are logged and analysed to measure Accuracy@10, ensuring that at least 75% of test cases return the correct result within the top ten answers. Identified errors and

[43] https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct. Accessed 26 Feb. 2025.
[44] https://w3id.org/OWLunit/ontology/. Accessed 26 Feb. 2025.

inconsistencies are then used to iteratively refine ontology structures, improving KG coverage and accuracy over time.

By automating these evaluations, the OWLUnit tool[45] facilitates continuous integration workflows, allowing ongoing ontology updates to be automatically tested. This ensures that the KG consistently meets predefined requirements and maintains its coverage and correctness as it evolves.

Rationale for Approach

We use different, complementary approaches to cover both bottom-up data-oriented knowledge extraction approaches and top-down expert-driven knowledge engineering approaches.

KPI3-a uses Retrieval-Augmented Generation (RAG) for systematic knowledge extraction from unstructured text, combining the strengths of retrieval-based and generative AI methods. This hybrid approach supports the creation of a robust knowledge graph that captures both explicit and implicit domain knowledge, critical for evidence-based decision-making.

1) **RAG significantly reduces hallucinations and enhances factual accuracy**, by grounding the generation process in relevant, curated knowledge bases. The methodology employs a comprehensive processing pipeline that includes entity extraction, type linking, triplet extraction, and the generation of detailed concept descriptions, so information is aligned with domain-specific requirements. Additionally, our approach **supports transparency** through preserving links to track the provenance of information.
2) **RAG's modular architecture allows flexible optimisation of system components**. Retrieval modules can be fine-tuned for specific domain knowledge, generation modules can be updated with advanced language models, and embedding models can be tailored to domain-specific contexts. This flexibility supports continuous system improvement without extensive redesigns.
3) **RAG supports enhanced scalability and maintenance**: new knowledge can be integrated into the retrieval index without retraining, the system can manage large volumes of medical texts, and updates to the knowledge base are instantly reflected in extractions. This makes RAG particularly suitable for dynamic, data-intensive fields like medical knowledge management, where maintaining accuracy, scalability, and adaptability is crucial.

KPI3-b, uses competency questions to assess knowledge graph (KG) coverage, grounded in the test-driven eXtreme Design (XD) methodology applied during ontology development. Competency questions serve as foundational inputs throughout the design, testing, and validation phases of ontology modules, ensuring alignment with the KG's ontological commitments and knowledge representation requirements. KPI assessment is conducted through competency question verification, which tests whether the ontology enables the translation of competency questions into corresponding SPARQL queries. These queries are then executed over the KG to verify if the expected answers are returned. This approach

[45] https://github.com/luigi-asprino/owl-unit. Accessed 26 Feb. 2025.

provides a practical, consolidated method for evaluating KG coverage, as it directly measures the system's ability to address specific, meaningful domain-related queries.

1) Competency questions help to **ensure alignment with domain requirements**, as competency questions capture the critical informational needs of the domain, and their successful execution demonstrates the presence of necessary entities, relationships, and data.
2) Converting competency questions into SPARQL queries **facilitates both structural and semantic validation**, testing the robustness of the ontology's design and the clarity of its semantic definitions.
3) This method **provides tangible evidence of coverage**, as successfully executed queries indicate that the KG accurately represents the domain's key aspects.
4) It **supports iterative improvement**, with failed competency questions highlighting gaps that guide refinements in the KG.
5) The approach also **ensures the KG remains relevant to real-world applications** because it derives questions from domain experts, making the assessment both practical and user-centric.

2.3.3. Limitations

KPI3a: Limitations include the scarcity of annotated datasets for comprehensive quantitative evaluation, the resource-intensive nature of high Top-K settings, the complexity of prompt engineering, and processing overhead in modular configurations. To address these challenges, recommendations include optimising Top-K settings for resource efficiency, exploring automated or adaptive prompt engineering techniques, and testing a broader range of commercial and open-source large language models to improve performance-cost trade-offs.

KPI3-b Limitations include heavy reliance on domain experts for defining and validating competency questions, which can be time-consuming and resource-intensive. Ambiguously defined competency questions can complicate SPARQL query formulation, while frequent changes in project requirements may increase workload in maintaining test cases. Errors in test case definitions can lead to misleading evaluation results, and the complexity of SPARQL may limit non-technical stakeholder involvement.

2.4. KPI 4 Evidence refinement

How much does the system engage users?

2.4.1. Relevant Objective

OBJ3: Case knowledge refinement: identify and enrich case-specific knowledge in support of evidence-based decision-making as a result of the collective inputs from humans and AI.

The goal of HACID is to support experts in the identification of relevant evidence in support of a specific case, resulting in the definition of structured knowledge that supports the hybrid collective intelligence in producing an optimal set of solutions.

2.4.2. Measurement Approach

Methodology for Measurement

Status: Planned

Metrics Used

The User Engagement Scale is a validated tool designed to measure user engagement in human-computer interaction contexts. Its latest version, the Refined User Engagement Scale, includes both a Long Form (30 items) and a Short Form (12 items).[46] [47] The User Engagement Scale-Short Form is selected for its efficiency and versatility across diverse applications, including mobile health evaluations.

It assesses four key dimensions of user engagement:

- Focused Attention
- Perceived Usability
- Aesthetic Appeal
- Reward Factor

This provides a comprehensive framework for evaluating user interactions with the HACID-DSS. A high engagement level suggests that the HACID-DSS effectively supports users in retrieving and interacting with system outputs.

- **Primary Metric:** User engagement measured via the Refined User Engagement Scale.
- **Measurement Criteria:** Evaluation of the four engagement dimensions (Focused Attention, Perceived Usability, Aesthetic Appeal, Reward Factor).
- **Target:** 80% user satisfaction, reflecting a high level of engagement and positive interaction with HACID-DSS's data retrieval capabilities.

[46] O'Brien, Heather L., et al. "A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form." *International Journal of Human-Computer Studies*, vol. 112, Apr. 2018, pp. 28–39. *ScienceDirect*, https://doi.org/10.1016/j.ijhcs.2018.01.004.

[47] Doherty, Kevin, and Gavin Doherty. "Engagement in HCI: Conception, Theory and Measurement." *ACM Comput. Surv.*, vol. 51, no. 5, Nov. 2018, p. 99:1-99:39. *ACM Digital Library*, https://doi.org/10.1145/3234149.

Data Sources

The User Engagement Scale will be applied through self-report surveys, where users will provide subjective evaluations of their engagement with the HACID-DSS.

Procedure

Participants engage with the HACID-DSS to solve a simple case, specifically using the knowledge graph visualisation and query system to explore functionalities. After completing the task, they fill out the User Engagement Scale questionnaire, selecting either the Long Form or Short Form based on feedback detail and time constraints.

Responses are processed by calculating the average scores for each engagement dimension, with negatively worded items reverse-scored.

Ideally, the User Engagement Scale is administered in two stages:

1. Establishing baseline engagement data, followed by system refinements based on feedback.
2. Reapplying the User Engagement Scale to measure improvements, aiming for a target mean score of four across all dimensions.

This structured approach ensures that HACID-DSS enhancements are data-driven and iteratively refined.

Rationale for Approach

Measuring the HACID-DSS system's ability to elicit user feedback is challenging due to the diversity of user interactions and feedback types. Actions like case enrichment, knowledge visualisation, and data queries vary widely, and neither the quantity nor type of interactions reliably reflects the quality of user feedback.

A detailed analysis of specific feedback types could provide deeper insights, but this would require complex experimental designs beyond the scope of this KPI, which focuses on general system engagement.

To address this, the User Engagement Scale is adopted as a standardised, validated framework for assessing user interaction and engagement. This scale is reliable across diverse contexts, allowing for objective evaluation of user engagement without relying solely on quantitative metrics or specific interaction types.

2.4.3. Limitations

The User Engagement Scale is effective for assessing general engagement; however, its broad focus may limit its ability to identify specific areas within the HACID-DSS and knowledge graph visualization system that require targeted improvements.

2.5. KPI 5 Collective accuracy

How often does the HACID-DSS include the correct solution as one of the top-ranked solutions in its proposed solution set?

2.5.1. Relevant Objective

OBJ4: Hybrid collective problem solving: harness the judgements of multiple experts by aggregating and expanding the set of provided solutions.

The goal of HACID is to reason over the available knowledge to optimally aggregate advice from multiple experts, expanding it on the basis of the knowledge available to provide improved decision performance (e.g., accuracy) while maintaining explainability.

2.5.2. Measurement Approach

Methodology for Measurement

Status: Complete

Metrics Used

Accurate decision-making is central to HACID-DSS, particularly in high-stakes applications such as medical diagnostics and open-ended expert problem-solving. The system aggregates individual expert inputs to improve accuracy beyond what any single expert could achieve alone. This KPI evaluates how well HACID-DSS enhances collective accuracy compared to individual decision-making. By optimally combining expert knowledge, the system aims to achieve a 40% improvement in accuracy, ensuring more reliable outcomes while maintaining explainability.

- Primary Metric: System's ability to identify correct solutions with high probability and reliability.
- Measurement Criteria:
    - Top-N Accuracy: Measures whether the correct solution appears within the top 1, 2, 3, or 5 ranked results, assessing how effectively the system prioritises accurate responses.
    - Mean Reciprocal Rank: A ranking-based metric from information retrieval, evaluating how highly the correct solution appears in the list of proposed answers.[48]
- **Target: 40% accuracy increase** compared to individual solutions, relative to the maximum possible improvement.

[48] Voorhees, Ellen M. "The TREC Question Answering Track." *Natural Language Engineering*, vol. 7, no. 4, Dec. 2001, pp. 361–78. *DOI.org (Crossref)*, https://doi.org/10.1017/S1351324901002789.

Data Sources

Data for the medical diagnostics use case are made available by Human Dx and have been collected through their online crowdsourcing platform. Our analyses use 2,133 medical cases and 40,762 differential diagnoses provided by medical professionals at various stages of their careers, including medical students (23%), interns (10%), residents (40%), fellows (1%), and attending physicians (26%). While specific studies utilise slightly varied subsets of this data due to differences in availability, all datasets maintain consistent core characteristics despite being collected at different times.

In addition to human-generated diagnoses, we evaluate the performance of large language models (LLMs) using the same set of case vignettes.[49] Five state-of-the-art LLMs are tested:

1. Anthropic Claude 3 Opus
2. Google Gemini Pro 1.0
3. Meta Llama 2 70B
4. Mistral Large
5. OpenAI GPT-4

The models are prompted to generate the five most probable diagnoses, ranked by their likelihood of being correct.

Procedure

A key challenge in aggregating independent diagnoses is determining whether different diagnoses refer to the same medical concept. This issue ranges from straightforward cases, such as spelling variations and typographical errors, to more complex instances, like identifying whether terms such as “OCD” and “Obsessive-Compulsive Disorder” refer to the same diagnosis. To address this, we use an automatic, reproducible, and scalable method that integrates semantic knowledge graphs and natural language processing with the SNOMED Clinical Terms ontology[50], enabling us to confidently determine when two diagnoses correspond to the same medical concept.

Next, we aggregate the standardised diagnoses and evaluate how often groups of varying sizes arrive at the correct solution using the defined accuracy metrics. We then compare the performance of collective accuracy to that of individual diagnoses, testing different aggregation strategies rooted in the wisdom-of-crowds approach. This evaluation follows the methodology outlined in Kurvers et al. (2023).[51]

---

[49] Zöller, N., Berger, J., Lin, I., Fu, N., Komarneni, J., Barabucci, G., Laskowski, K., Shia, V., Harack, B., Chu, E. A., Trianni, V., Kurvers, R. H. J. M., & Herzog, S. M. (2024, June 21). Human-AI collectives produce the most accurate differential diagnoses. arXiv.Org. https://arxiv.org/abs/2406.14981v1

[50] Kurvers, Ralf H. J. M., et al. “Automating Hybrid Collective Intelligence in Open-Ended Medical Diagnostics.” *Proceedings of the National Academy of Sciences*, vol. 120, no. 34, Aug. 2023, p. e2221473120. *DOI.org (Crossref)*, https://doi.org/10.1073/pnas.2221473120.

[51] Kurvers, Ralf H. J. M., et al. “Automating Hybrid Collective Intelligence in Open-Ended Medical Diagnostics.” *Proceedings of the National Academy of Sciences*, vol. 120, no. 34, Aug. 2023, p. e2221473120. *DOI.org (Crossref)*, https://doi.org/10.1073/pnas.2221473120.

Rationale for Approach

The HACID approach aims to substantially improve decision performance (i.e., accuracy) when aggregating the judgements of multiple independent experts in open-ended decision-making contexts. The goal is to ensure that high-quality solutions consistently rank at the top of aggregated outputs. Identifying high-quality decisions is critical in expert domains, particularly where uncertainty and vast open-ended solution spaces complicate the decision-making process. The complexity of these environments makes it essential to implement robust methods for reliably and effectively identifying promising options.

A key strength of HACID is its ability to address the challenge of combining decisions from independent raters, particularly in open-ended domains where different responses may refer to the same concept. By integrating semantic knowledge graphs and natural language processing with a standardised medical ontology, HACID ensures conceptually similar diagnoses are correctly identified and aggregated, improving the quality of collective decision-making. Additionally, leveraging a hybrid approach that combines structured aggregation techniques with expert input enhances explainability, ensuring results remain interpretable and actionable.

2.5.3. Limitations

The reliance on data from the Human Dx platform may restrict generalisability, given that the dataset may not capture the full spectrum of medical cases, including rare or highly complex scenarios. Additionally, since the data is based on clinical vignettes rather than real-world patient interactions, the extent to which these findings apply to actual clinical practice remains uncertain.[52]

[52] Peabody, John W., et al. “Measuring the Quality of Physician Practice by Using Clinical Vignettes: A Prospective Validation Study.” *Annals of Internal Medicine*, vol. 141, no. 10, Nov. 2004, pp. 771–80. *PubMed*, https://doi.org/10.7326/0003-4819-141-10-200411160-00008.

2.6. KPI 6 Collective effectiveness

How effective is the HACID-DSS in providing actionable outcomes?

2.6.1. Relevant Objective

OBJ4: Hybrid collective problem solving: harness the judgements of multiple experts by aggregating and expanding the set of provided solutions.

The goal of HACID is to reason over the available knowledge to optimally aggregate advice from multiple experts, expanding it on the basis of the knowledge available to provide improved decision performance (e.g., accuracy) while maintaining explainability.

2.6.2. Measurement Approach

Methodology for Measurement

Status: Planned

Metrics Used

The evaluation of KPI6 focuses on measuring the effectiveness of HACID in aggregating expert knowledge and improving decision-making performance while maintaining explainability. Given the open-ended nature of problems in domains such as climate services, where no universal ground truth exists, this KPI relies on representative indicators to assess the quality and actionability of solutions generated by the system.

Evaluation divides stakeholders into two groups: 1) Case Creators, who generate cases (with their respective benchmark solutions) and assess how collective solutions compare to individual inputs; and 2) Case Solvers, who evaluate the usefulness of aggregated solutions.

- **Primary Metric:** Effectiveness of HACID-DSS in aggregating expert knowledge and improving decision-making performance.
- **Measurement Criteria:**
    - Case Creators
        - Quality Improvement: Measures whether collective solutions are rated higher than individual solutions in terms of accuracy and completeness.
        - Benchmark Revisions: Tracks how often case creators update their ex-ante benchmark after reviewing collective solutions, reflecting shifts in expert understanding (i.e., tracking the appearance of novel insights).
    - Case Solvers
        - Perceived Usefulness: Proportion of case solvers who rate the aggregated solutions as beneficial for decision-making.
- **Target:** ≥ 70% of practitioners in climate services must perceive collective solutions as useful.

Data Sources

Two crowdsourcing studies provide the empirical basis for evaluating KPI6:

- **CS-CROWD (Baseline Crowdsourcing Study):** Establishes initial benchmarks and evaluates the effectiveness of collective solutions through structured case development, solution elicitation, aggregation, and assessment.
- **CS-EXPERIMENT (Follow-up Crowdsourcing Experiment):** Builds upon CS-CROWD by testing interventions designed to improve the quality of collective solutions.

Procedure

To evaluate KPI6, HACID is developing an empirical protocol based on CS-CROWD, designed to assess the collective effectiveness of decision support within climate services. The process includes:

1. **Case Development**: Climate services experts from the Met Office create climate-related cases and provide an initial benchmark solution (ex-ante benchmark).
2. **Solution Elicitation**: Climate science experts contribute solutions to these cases via a dedicated HACID interface.
3. **Solution Aggregation**: Individual solutions are automatically aggregated to form collective solutions.
4. **Assessment of Solutions**: Case creators evaluate the quality of individual and aggregated solutions while remaining blinded to their source.
5. **Benchmark Revision**: Based on the solutions received, case creators may revise their initial benchmark, creating an ex-post benchmark.
6. **Solver Feedback**: Case solvers assess the usefulness of collective solutions, offering additional qualitative insights.
7. **Post benchmark evaluation:** A complimentary evaluation performed by case solvers, to understand how good the benchmark is considered by community experts

Building upon this baseline, **CS-EXPERIMENT** tests specific interventions aimed at enhancing the quality of aggregated solutions. This second phase seeks to refine the HACID framework by introducing mechanisms to improve decision accuracy and usability while maintaining fairness and transparency in expert aggregation.

Rationale for Approach

Evaluating HACID's collective effectiveness is particularly complex in domains like climate services, where no single ground truth exists. Traditional evaluation methods are inadequate, necessitating a reliance on representative indicators to assess solution quality and actionability.

The approach ensures structured aggregation of expert input, improving the clarity and reliability of decision-making. The evaluation framework is tailored to climate services, incorporating expert knowledge to validate the effectiveness of HACID's decision-support processes in real-world scenarios. By integrating structured case development and

systematic solution aggregation, HACID aims to enhance both accuracy and applicability, ensuring solutions are interpretable and actionable.

2.6.3. Limitations

The primary limitation of KPI 6's evaluation approach remains the scarcity of domain experts, which constrains the number of climate-service cases developed and the volume of expert solutions. Additionally, assessing solutions in the absence of a definitive ground truth remains inherently challenging.

2.7. KPI 7 Collective efficiency

How efficient is the HACID-DSS in terms of decision costs?

2.7.1. Relevant Objective

OBJ4: Hybrid collective problem solving: harness the judgements of multiple experts by aggregating and expanding the set of provided solutions.

The goal of HACID is to reason over the available knowledge to optimally aggregate advice from multiple experts, expanding it on the basis of the knowledge available to provide improved decision performance (e.g., accuracy) while maintaining explainability.

2.7.2. Measurement Approach

Methodology for Measurement

Status: Completed

Metrics Used

The primary metric for assessing the efficiency of a target aggregation method (see KPI-8) measures the number of raters required to achieve certain levels of accuracy from a baseline (established by aggregating decisions from randomly selected individuals). For instance, the accuracy achieved by a group of five randomly selected individuals serves as a benchmark, and the goal is to reach similar accuracy levels with fewer raters (e.g., by selecting individuals with high performance on other cases; or selecting a smaller number of individuals based on their decision similarity), thereby reducing costs without sacrificing performance.

Since accuracy can be evaluated in different ways, multiple metrics are used to assess performance at varying levels of granularity.

- **Primary Metric:** Number of raters required to maintain target accuracy while minimising decision-making costs.
- **Measurement Criteria:**
    - Baseline Comparison: Reduction in required raters while maintaining accuracy.
    - Top-N Accuracy: Evaluates whether the correct diagnosis appears within the top 1, 2, 3, or 5 ranked solutions.
    - Mean Reciprocal Rank: Assesses how high the correct solution is ranked in a list of possible answers.[53]
- **Target:** Significant reduction in costs compared to baseline aggregation methods, without compromising accuracy.

[53] Voorhees, E. M. (1999, November). The trec-8 question answering track report. In *Trec* (Vol. 99, pp. 77-82). https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=151495

Data Sources

Data for this KPI are made available by Human Dx (see KPI-5: 'Data Sources').

Procedure

The analysis follows a structured approach:

1. **Data Standardisation** – Diagnoses are aligned using **semantic knowledge graphs, natural language processing, and the SNOMED Clinical Terms ontology** to ensure consistent identification of medical conditions (as outlined in KPI 5).
2. **Aggregation and Benchmarking** – The diagnostic accuracy of collectives of varying sizes are computed using computer simulations. The resulting accuracy metrics (Top-N accuracy, mean reciprocal rank) then allow assessing how many expert inputs are required for the target aggregation method to maintain diagnostic accuracy (relative to the baseline aggregation method) while minimising resource costs.
3. **Comparison of Human and Hybrid Crowds** – Expert-only decision-making is benchmarked against **hybrid crowds**, where expert input is supplemented with large language model (LLM)-generated predictions.

The efficiency gains from hybrid models were compared against human-only crowds to assess the potential of LLMs in optimising diagnostic decision-making.

Rationale for Approach

The HACID approach aims to enhance accuracy while ensuring cost-effective decision-making, particularly in expert-driven fields such as medicine. While collective aggregation improves performance, it also increases decision costs. By improving the selection of expert inputs and integrating hybrid models, HACID seeks to balance accuracy and efficiency.

This study extends existing research on decision similarity heuristics, demonstrating that high-performing individuals are identifiable in open-ended medical decision-making, reducing the need for large expert panels.[54] [55] Additionally, the comparative evaluation of human-only and hybrid crowds highlights significant efficiency gains, reinforcing the potential of AI-assisted expert aggregation in complex problem-solving domains.[56]

---

[54] Zöller, N., Herzog, S. M., & Kurvers, R. H. J. M. (2023). Boosting collective intelligence in medical diagnostics: Leveraging decision similarity as a predictor of accuracy when answers are open-ended rankings. HCOMP-CI 2023 Works-in-Progress and Demonstrations. https://www.humancomputation.com/assets/wips_demos/CI-23_paper_1055.pdf

[55] Kurvers, R. H. J. M., Herzog, S. M., Hertwig, R., Krause, J., Moussaid, M., Argenziano, G., Zalaudek, I., Carney, P. A., & Wolf, M. (2019). How to detect high-performing individuals and groups: Decision similarity predicts accuracy. Science Advances, 5(11), eaaw9011. https://doi.org/10.1126/sciadv.aaw9011

[56] Zöller, N., Berger, J., Lin, I., Fu, N., Komarneni, J., Barabucci, G., Laskowski, K., Shia, V., Harack, B., Chu, E. A., Trianni, V., Kurvers, R. H. J. M., & Herzog, S. M. (2024, June 21). Human-AI collectives produce the most accurate differential diagnoses. arXiv.Org. https://arxiv.org/abs/2406.14981v1

2.7.3. Limitations

The reliance on data from the Human Dx platform may restrict generalisability, given that the dataset may not capture the full spectrum of medical cases, including rare or highly complex scenarios. Additionally, since the data is based on clinical vignettes rather than real-world patient interactions, the extent to which these findings apply to actual clinical practice remains uncertain (see also KPI-5).[57]

[57] Peabody, J. W., Luck, J., Glassman, P., Jain, S., Hansen, J., Spell, M., & Lee, M. (2004). Measuring the quality of physician practice by using clinical vignettes: a prospective validation study. *Annals of internal medicine*, *141*(10), 771-780.. https://doi.org/10.7326/0003-4819-141-10-200411160-00008

2.8. KPI 8 Metadata boosts

How do individual confidence, response times, written justifications, and expertise affect collective solutions?

2.8.1. Relevant Objective

OBJ5: Advice reinforcement in hybrid collective problem solving: develop methods to improve the way in which experts provide their judgements and solutions—in isolation and in interaction with other experts—to support advanced aggregation methods.

The goal of HACID is to determine context-specific approaches to tap into individual knowledge by (i) eliciting confidence and justifications of proposed solutions, (ii) weighing individual expertise on the basis of the history of interactions and (iii) exploiting social influence by building suitable interaction networks among experts that can promote better decisions by overcoming individual and social biases.

2.8.2. Measurement Approach

Methodology for Measurement

Status: In progress

Metrics Used

Metadata such as confidence levels, response times, written justifications, and past performance can significantly influence the effectiveness of collective decision-making. This KPI assesses how metadata-driven aggregation strategies enhance the system's ability to identify the correct solution with high probability and reliability. By leveraging metadata, HACID-DSS aims to improve the accuracy of collective solutions beyond what individual experts can achieve alone.

For the medical diagnostics use case, the goal is to achieve a 50% improvement in accuracy compared to individual solutions, relative to the maximum possible accuracy gain. For example, if individual accuracy is 50%, the target would be an absolute increase of 25% in accuracy. Because accuracy can be expressed in multiple ways in open-ended domains, various metrics are used to evaluate performance at different levels of granularity.

- **Primary Metric:** The collective's ability to identify the correct solution with high probability and reliability.
- **Measurement Criteria:**
    - Top-N Accuracy: Evaluates whether the correct solution appears within the top 1, 2, 3, or 5 ranked results.
    - Mean Reciprocal Rank: Assesses how high the correct solution is ranked within the proposed answer list.
    - Metadata Influence: Measures how confidence levels, response times, and written justifications impact collective accuracy (see KPI-5) and collective efficiency (see KPI-7).

- **Target:** 50% accuracy increase compared to individual solutions, relative to the maximum possible improvement (i.e., 50% * [100% - average individual accuracy]).

Data Sources

The study draws on data acquired by Human Dx through their crowdsourcing platform. The dataset includes 30 medical cases, with at least 20 responses per case, capturing confidence ratings alongside differential diagnoses. Additionally, a larger existing Human Dx dataset (50 cases, with 200 solves per case) is used to analyse response times and prior performance, ensuring comprehensive metadata analysis.

Procedure

To ensure consistency, data collection is standardised throughout the experiment. When eliciting diagnosticians' confidence levels, we employ phrasing aligned with established practices in cognitive psychology to minimise potential misunderstandings of the confidence scale. This approach ensures clarity and reliability in the confidence data collected, supporting the robustness of subsequent analyses.

See KPI-5 and KPI-7 for the procedure to evaluate collective accuracy and collective efficiency, respectively.

Rationale for Approach

HACID leverages metadata to improve decision aggregation[58] [59], refining expert-based solutions by incorporating confidence assessments,[60] [61] [62] [63] response speeds, and justification strength.[64] Traditional approaches often overlook these factors, yet they play a critical role in evaluating the reliability of expert inputs.

---

[58] Collins, R. N., Mandel, D. R., & Budescu, D. V. (2023). Performance-weighted aggregation: Ferreting out wisdom within the crowd. In M. Seifert (Ed.), Judgment in Predictive Analytics (pp. 185–214). Springer International Publishing. https://doi.org/10.1007/978-3-031-30085-1_7

[59] Atanasov, P., & Himmelstein, M. (2023). Talent spotting in crowd prediction. In M. Seifert (Ed.), Judgment in Predictive Analytics (pp. 135–184). Springer International Publishing. https://doi.org/10.1007/978-3-031-30085-1_6

[60] Meyen, S., Sigg, D. M. B., Luxburg, U. von, & Franz, V. H. (2021). Group decisions based on confidence weighted majority voting. Cognitive Research: Principles and Implications, 6(1), 18. https://doi.org/10.1186/s41235-021-00279-0

[61] Meyen, S., Sigg, D. M. B., Luxburg, U. von, & Franz, V. H. (2021). Group decisions based on confidence weighted majority voting. Cognitive Research: Principles and Implications, 6(1), 18. https://doi.org/10.1186/s41235-021-00279-0

[62] Bennett, S. T., Benjamin, A. S., Mistry, P. K., & Steyvers, M. (2018). Making a wiser crowd: Benefits of individual metacognitive control on crowd performance. Computational Brain & Behavior, 1(1), 90–99. https://doi.org/10.1007/s42113-018-0006-4

[63] Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. Journal of Personality and Social Psychology, 107(2), 276–299. https://doi.org/10.1037/a0036677

[64] Kotamraju, S., & Blanco, E. (2021). Written justifications are key to aggregate crowdsourced forecasts. arXiv:2109.07017 [Cs]. http://arxiv.org/abs/2109.07017

This method represents an innovative enhancement in collective decision-making, particularly in open-ended contexts where defining a universal ground truth is difficult. By integrating metadata-driven aggregation, HACID seeks to identify high-quality contributions more accurately and enhance decision explainability.

2.8.3. Limitations

The reliance on data from the Human Dx platform may restrict generalisability, given that the dataset may not capture the full spectrum of medical cases, including rare or highly complex scenarios. Additionally, since the data is based on clinical vignettes rather than real-world patient interactions, the extent to which these findings apply to actual clinical practice remains uncertain (see also KPI-5).[65]

[65] Peabody, J. W., Luck, J., Glassman, P., Jain, S., Hansen, J., Spell, M., & Lee, M. (2004). Measuring the quality of physician practice by using clinical vignettes: a prospective validation study. *Annals of internal medicine*, *141*(10), 771-780. https://doi.org/10.7326/0003-4819-141-10-200411160-00008

2.9. KPI 9 Social information boosts

How does social information affect collective solutions?

2.9.1. Relevant Objective

OBJ5: Advice reinforcement in hybrid collective problem solving: develop methods to improve the way in which experts provide their judgements and solutions—in isolation and in interaction with other experts—to support advanced aggregation methods.

The goal of HACID is to determine context-specific approaches to tap into individual knowledge by (i) eliciting confidence and justifications of proposed solutions, (ii) weighing individual expertise on the basis of the history of interactions and (iii) exploiting social influence by building suitable interaction networks among experts that can promote better decisions by overcoming individual and social biases.

2.9.2. Measurement Approach

Methodology for Measurement

Status: in progress

Metrics Used

Social information—such as expert exposure to peer decisions or shared reasoning processes—can influence decision accuracy in collective intelligence systems.[66] This KPI evaluates whether providing social cues improves diagnostic accuracy without introducing bias or inefficiency. The system's ability to identify the correct solution with high probability and reliability is assessed across multiple levels of accuracy measurement.

Since accuracy in open-ended decision-making can be measured in various ways, multiple metrics are employed to capture performance at different granularities.

- **Primary Metric:** System's ability to identify the correct solution with high probability and reliability in the presence of social information.
- **Measurement Criteria:**
    - Top-N Accuracy: Evaluates whether the correct diagnosis appears within the top 1, 2, 3, or 5 ranked solutions.
    - Mean Reciprocal Rank: Measures how high the correct solution is ranked within the proposed answer set.[67]
    - Social Information Exposure: Assesses the effect of expert exposure to peer decisions on collective accuracy (see KPI-5) and collective efficiency (see KPI-7).

---

[66] Kameda, T., Toyokawa, W., & Tindale, R. S. (2022). Information aggregation and collective intelligence beyond the wisdom of crowds. Nature Reviews Psychology, 1(6), 345–357. https://doi.org/10.1038/s44159-022-00054-y .

[67] "The TREC-8 Question Answering Track Report", *NIST,* https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=15149. Accessed 26 Feb.2025

- **Target:** Non-negative impact on accuracy compared to the baseline, with little to no loss in efficiency.

Data Sources

We have devised two experimental studies, referred to as CLINIFLOW and MDT. Data are collected by Human Dx through their online crowdsourcing platform. Each case is presented as a vignette that closely simulates real-world diagnostic scenarios, including patient symptoms, medical histories, and clinical test results. This dataset ensures that evaluations are based on realistic, expert-driven decision-making processes, making it a valuable resource for assessing HACID's diagnostic decision-support capabilities.

Procedure

The two studies collect solutions to medical cases placing users in different experimental conditions. In both studies, the metrics discussed above are computed for the experimental condition and compared to a baseline, to assess how the administered treatment influences decision making. In the following, we describe separately the experimental conditions devised for both planned studies.

CLINIFLOW

The CLINIFLOW study explores how human problem solvers' performance in open-ended general medical diagnostics can be enhanced by introducing advice from human, AI, or mixed human-AI advisors. A key focus is on the timing of advice, comparing its impact when provided early versus late in the problem-solving process. *Figure 2.2* illustrates the high-level experimental design, while *Figure 2.3* shows examples of the Human Diagnosis Project (HDx) platform interface as experienced by participants across different experimental conditions

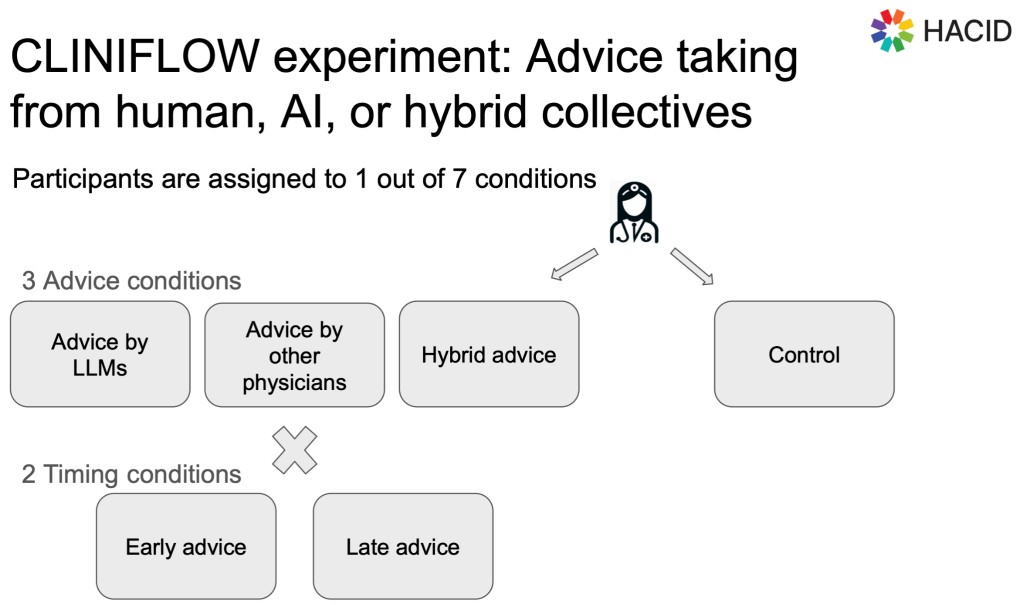

Figure 2.2: Overview of the CLINIFLOW experimental design

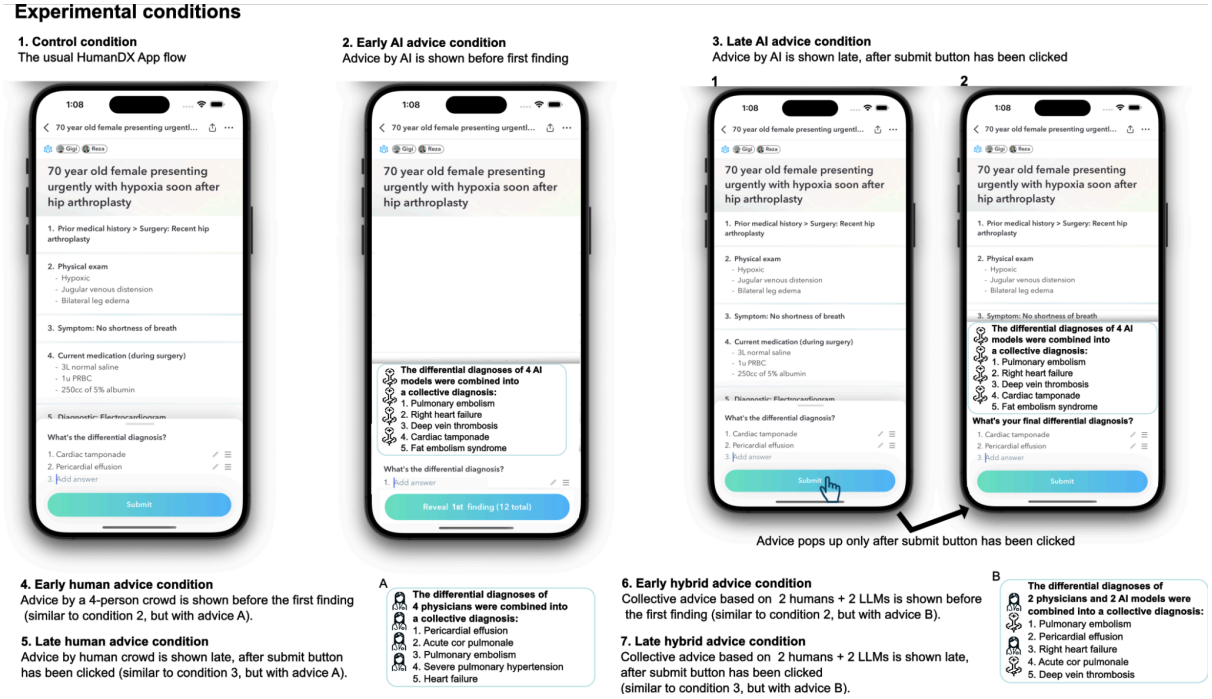

FIgure 2.3: Examples of HDx platform interface as it appears across experimental conditions.

MDT (Multi-Disciplinary Teams)

The MDT study investigates the effectiveness of "wisdom of small, deliberative crowds"[68] [69] in open-ended medical diagnostics. Specifically, it examines:

1. Whether aggregating the outputs of multiple small, interactive teams—each producing a group differential diagnosis after deliberation—yields higher diagnostic accuracy than (i) decisions made by individual experts and (ii) simple aggregations of individual solutions.
2. The role of diversity (e.g., gender diversity as highlighted in Nielsen et al., 2017[70]) in influencing team deliberation processes and the accuracy of aggregated outcomes.

A critical challenge in both studies is accurately aggregating and comparing independent diagnoses in open-ended medical contexts (see also previous KPIs).

Rationale for Approach

The HACID approach leverages social influence to improve decision accuracy by strategically integrating expert insights into decision-support systems. Traditional aggregation

[68] Navajas, Joaquin, et al. "Aggregated Knowledge from a Small Number of Debates Outperforms the Wisdom of Large Crowds." *Nature Human Behaviour*, vol. 2, no. 2, Feb. 2018, pp. 126–32. *www.nature.com*, https://doi.org/10.1038/s41562-017-0273-4.

[69] Dezecache, Guillaume, et al. "Democratic Forecast: Small Groups Predict the Future Better than Individuals and Crowds." *Journal of Experimental Psychology: Applied*, vol. 28, no. 3, 2022, pp. 525–37. *APA PsycNet*, https://doi.org/10.1037/xap0000424.

[70] Nielsen, Mathias Wullum, et al. "Gender Diversity Leads to Better Science." *Proceedings of the National Academy of Sciences*, vol. 114, no. 8, Feb. 2017, pp. 1740–42. *DOI.org (Crossref)*, https://doi.org/10.1073/pnas.1700616114.

methods often overlook the role of peer advice, deliberation, and hybrid human-AI collaboration, which can provide substantial accuracy gains when carefully structured. However, social feedback could also lead to herding that undermines decision accuracy. Hence, we are particularly interested in interventions that do not reduce decision accuracy. The observation of a significant accuracy gain would be a plus.

KPI 9 draws on established research in group decision-making, advice-taking, and crowdsourcing, applying these principles to open-ended medical diagnostics. The CLINIFLOW study represents an innovative approach, as it is the first known research to examine hybrid human-AI advice-taking in such domains. Additionally, the MDT study investigates the wisdom of structured, deliberative small teams, offering a new avenue for optimising collective decision-making.

2.9.3. Limitations

This study faces practical constraints due to its implementation within a live platform (Human Dx), which limits the volume of cases and responses compared to controlled lab experiments. Additionally, the dataset may not fully represent rare or highly complex medical cases, which could impact general applicability.

2.10. KPI 10 Process evaluation

How well do the evaluation methods capture the criteria that matter to decision makers?

2.10.1. Relevant Objective

OBJ6: Evaluation: develop an evaluation framework for decision support that covers diverse aspects relevant for the application context.

The goal of HACID is to evaluate the usage of the DSS in practice, identifying the potential for improvement in supporting complex decisions, and the effects it can have on human engagement and decision making, also with reference to interpretability, explainability and trustworthiness.

2.10.2. Measurement Approach

Methodology for Measurement

Status: In progress

Metrics Used

This KPI evaluates whether HACID-DSS evaluation methods accurately reflect stakeholder needs and decision-making priorities. Expert validation is gathered through structured surveys and participatory workshops, assessing perceptions of inclusivity, clarity, transparency, and appropriateness in the evaluation process.

- **Primary Metric:** Validation of defined evaluation criteria by domain experts.
- **Measurement Criteria:**
    - Structured survey responses assessing stakeholder perceptions of inclusion, clarity, transparency, and appropriateness.
    - Expert feedback from values elicitation and risk assessment workshops.
- **Target:** ≥ 80% of participating experts must provide a positive or very positive assessment of the evaluation framework.

Data Sources

Data are collected from two participatory workshops:

1. Values Elicitation Workshop (Completed – Climate Domain) – Captures stakeholder-defined values for decision-support systems.
2. Risk Assessment Workshop (Planned – Medical Domain) – Assesses perceived risks of AI-driven decision support in healthcare, involving both medical professionals and the general public.

Survey responses serve as a validation mechanism, ensuring the evaluation framework aligns with real-world stakeholder concerns (Survey items listed in Appendix 1).

.

Procedure

1. Values Elicitation Workshop (Completed - Climate Domain)

This workshop engages climate scientists in a facilitated deliberation to explore what they value in HACID-style decision-support systems. Participants:

- Identify key values through guided discussion.
- Rank surfaced values via a pairwise comparison task, resulting in a prioritised list.
- Tag elements of hypothetical system-generated solutions to indicate which aspects increase or decrease trust, and provide justifications (linked to their values).

These outputs are synthesised into recommendations for the development team, ensuring stakeholder values **inform** system design. Participants then complete the survey items outlined earlier, measuring their satisfaction with how their contributions were captured.

2. Risk Assessment Workshop (Planned - Medical Domain)

This workshop focuses on eliciting perceived risks of AI-driven decision support in healthcare by bringing together both medical professionals and members of the general public. Specifically, it aims to:

- Compare risk perceptions between expert and non-expert stakeholders.
- Identify risk mitigation strategies.
- Assess how perceived risks vary for Hybrid-CI systems vs. fully automated or non-AI systems.

The structured deliberative format encourages discussion, leading to a list of stakeholder-defined risks. As with the values elicitation workshop, findings are synthesised into recommendations for the HACID development team, and participants complete the survey items to assess their satisfaction with how well their input is captured.

Rationale for Approach

Ensuring that HACID-DSS evaluation criteria align with stakeholder values and adequately address perceived risks is essential due to its hybrid human-AI nature. This integration ensures the system is not only technically effective but also trusted, contextually relevant, responsive to real-world needs, and safe. By embedding stakeholder validation into the evaluation framework—across values integration and risk assessment—HACID fosters technical excellence, ethical integrity, and a comprehensive approach to stakeholder safety.

Research[71] highlights how moral-reasoning frameworks such as the Veil of Ignorance—a decision-making principle that promotes impartiality by asking individuals to design policies

[71] Weidinger, Laura, et al. “Using the Veil of Ignorance to Align AI Systems with Principles of Justice.” *Proceedings of the National Academy of Sciences*, vol. 120, no. 18, May 2023, p. e2213709120. *DOI.org (Crossref)*, https://doi.org/10.1073/pnas.2213709120.

without knowledge of their own societal position—can help elicit principles of fairness and objectivity for AI systems. Similarly, alignment literature emphasises that interpretability, controllability, and ethicality are key to stakeholder trust and system relevance.

Risk assessment also plays a critical role in AI evaluation, ensuring systems align with stakeholder values while mitigating potential harms. AI risks span multiple domains, from misinformation and bias to security vulnerabilities and human rights concerns.[72] Engaging stakeholders in risk assessment workshops broadens understanding of these risks, fostering transparency and accountability in AI deployment. By integrating participatory risk assessment into HACID's evaluation framework, we ensure that both technical and societal risks are systematically identified and addressed, strengthening trust and safety in hybrid decision-support systems.

2.10.3. Limitations

The primary challenge remains ensuring diversity in stakeholder engagement, as expert-driven discussions tend to overlook broader societal concerns. Additionally, translating stakeholder values into concrete evaluation criteria and system design recommendations remains complex. The time-intensive nature of workshops poses scalability challenges, potentially constraining broader participation.

[72] Abercrombie, Gavin, et al. *A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms*. arXiv:2407.01294, arXiv, 9 Nov. 2024. *arXiv.org*, https://doi.org/10.48550/arXiv.2407.01294.

2.11. KPI 11 Output evaluation

To what extent are the HACID-DSS outputs aligned with stakeholder values?

3.11.1. Relevant Objective

OBJ6: Evaluation—Develop an evaluation framework for decision support that covers diverse aspects relevant to the application context.

The goal of HACID is to assess how effectively the DSS supports complex decision-making, evaluating its impact on human engagement, interpretability, explainability, and trustworthiness.

2.11.2. Measurement Approach

Methodology for Measurement

Status: In progress

Metrics Used

For HACID-DSS to be effective and widely accepted, its evaluation framework must align with stakeholder values and concerns across different domains. This KPI assesses how well HACID's outputs reflect stakeholder-defined priorities, ensuring that system development remains transparent, trustworthy, and aligned with expert expectations.

To establish meaningful evaluation criteria, stakeholder-defined values and risks are elicited through participatory processes and then integrated into structured evaluation frameworks. In the climate domain, trust factors such as transparency, uncertainty awareness, and bias mitigation are prioritised to guide prototype evaluation. In the medical domain, a risk assessment workshop is used to surface concerns about AI-driven decision support, forming a complementary evaluation framework.

Evaluation is conducted through structured survey responses and facilitated deliberations, allowing experts to assess HACID-DSS outputs against predefined value- and risk-based criteria.

- **Primary Metric:** Validation of evaluation criteria by domain experts.
- **Measurement Criteria:**
    - Survey-based assessment: Likert-scale survey items measuring perceived alignment of evaluation criteria with stakeholder-defined values and risks.
    - Expert feedback (qualitative) from structured discussions where stakeholders critically evaluate HACID-DSS using predefined evaluation prompts.
- **Target:** ≥ 80% of participating experts must provide a positive or very positive assessment of how well the evaluation framework captures their values and concerns.

Data Sources

Key data sources include:

- **Findings from a values elicitation workshop** (climate scientists) – providing a structured list of stakeholder-ranked values.
- **Findings from the medical risk assessment session** (planned) – identifies key stakeholder-perceived risks.
- **Survey responses** collected after workshops and prototype evaluations, providing direct feedback on alignment with values and risks.
- **Qualitative insights from structured deliberations**, capturing stakeholder reasoning behind their assessments of HACID's outputs.

Procedure

The evaluation of HACID's alignment with stakeholder values and perceived risks occurs through a two-stage participatory process:

1. Defining the Evaluation Framework (directly related to work carried out for KPI 10)
    - Climate domain: Values elicitation workshop – surfaces, prioritises, and ranks stakeholder-defined values.
    - Medical domain: Risk assessment workshop – identifies and prioritises stakeholder-perceived risks.
    - Workshop outputs are synthesised into structured evaluation criteria to be applied in prototype evaluation sessions.
2. Prototype Evaluation *(conducted in both climate and medical domains)*
    - Stakeholder groups (e.g., climate scientists, policymakers, general public) interact with HACID-generated outputs and evaluate them against the predefined values- and risk-based criteria.
    - Survey responses measure the degree to which participants believe HACID's outputs align with stakeholder priorities.
    - Facilitated discussions allow for qualitative assessment of system performance, highlighting areas for refinement.

This iterative, stakeholder-driven approach ensures that HACID is evaluated in ways that reflect real-world concerns, priorities, and expectations, ultimately strengthening its safety, credibility, and trustworthiness.

Rationale for Approach

This KPI evaluates how well HACID-generated outputs reflect the values, priorities, and concerns of decision-makers, embedding stakeholder-defined evaluation criteria into the assessment process. By integrating insights from values elicitation and risk assessment workshops (see KPI 10), we ensure a structured, participatory approach to evaluating HACID's effectiveness.

This approach enhances the legitimacy of HACID's outputs while strengthening stakeholder trust. Research on value alignment in AI highlights the importance of stakeholder-driven

assessments to ensure systems meet real-world expectations.[73] [74] While the relationship between such assessments and trust or interpretability is complex,[75] community-centered deliberation can elicit diverse perspectives, particularly from underrepresented groups[76]; and insights from work on deliberative AI governance show that transparent, stakeholder-driven assessments contribute to accountability and adaptability in AI systems.[77] [78] [79] [80]. Embedding these principles ensures HACID is not only technically effective but also ethically grounded and responsive to stakeholder needs.

In addition to value alignment, assessing risk perceptions in HACID's outputs is crucial. AI-generated decisions may introduce biases, uncertainty, or unintended consequences, requiring structured risk assessment methods. Stakeholder engagement in facilitated deliberations allows for a systematic evaluation of risks, ensuring that HACID outputs meet both technical and societal expectations. By embedding risk-based evaluation criteria, HACID strengthens the integrity and reliability of its decision-support outputs, fostering greater stakeholder confidence in its real-world applicability.

2.11.3. Limitations

Scalability is a key limitation—while participatory evaluation ensures depth, it is resource-intensive and difficult to generalise across large populations. Additionally, stakeholder diversity and power dynamics may shape evaluations, as expert-led discussions could marginalise non-expert perspectives. Translating qualitative insights into quantifiable evaluation metrics is also complex, requiring careful methodological design to ensure consistency across domains.

---

[73] "AI Value Alignment: Guiding Artificial Intelligence Towards Shared Human Goals.", *Global Future Council,* https://www3.weforum.org/docs/WEF_AI_Value_Alignment_2024.pdf. Accessed 26 Feb. 2025

[74] Bergman, Stevie, et al. "STELA: A Community-Centred Approach to Norm Elicitation for AI Alignment." *Scientific Reports*, vol. 14, no. 1, Mar. 2024, p. 6616. *www.nature.com*, https://doi.org/10.1038/s41598-024-56648-4.

[75] Weidinger, Laura, et al. "Using the Veil of Ignorance to Align AI Systems with Principles of Justice." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 120, no. 18, p. e2213709120. *PubMed Central*, https://doi.org/10.1073/pnas.2213709120.

[76] "Participatory AI for Humanitarian Innovation: A Briefing Paper." *Nesta*, https://www.nesta.org.uk/report/participatory-ai-humanitarian-innovation-briefing-paper/. Accessed 25 Feb. 2025.

[77] Reuel, Anka, and Trond Arne Undheim. *Generative AI Needs Adaptive Governance*. arXiv:2406.04554, arXiv, 6 June 2024. *arXiv.org*, https://doi.org/10.48550/arXiv.2406.04554.

[78] "Building AccountableAI Programs: Mapping Emerging Best Practices to the CIPL Accountability Framework.", *CIPL,* https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_building_accountable_ai_programs_23_feb_2024.pdf. Accessed 26 Feb. 2025

[79] "AI Value Alignment: Guiding Artificial Intelligence Towards Shared Human Goals.", *Global Future Council,* https://www3.weforum.org/docs/WEF_AI_Value_Alignment_2024.pdf. Accessed 26 Feb. 2025

[80] Shen, Hua, et al. *ValueCompass: A Framework of Fundamental Values for Human-AI Alignment*. arXiv:2409.09586, arXiv, 15 Sept. 2024. *arXiv.org*, https://doi.org/10.48550/arXiv.2409.09586.

2.12. KPI 12 Participatory design:

Have we introduced new participatory approaches to AI development?

2.12.1. Relevant Objective

OBJ7: Participatory AI—deploy a participatory approach to the development of hybrid collective intelligence exploiting human expertise and AI.

The goal of HACID is to make the HACID-DSS deployable across diverse application sectors through a participatory AI approach, showcasing the potential for reusing concepts developed within the project across different domains.

2.12.2. Measurement Approach

Methodology for Measurement

Status: In progress

Metrics Used

This KPI measures the extent to which participatory interventions—including workshops, policy discussions, and collaborative design activities—are integrated into system development. Stakeholder engagement at multiple stages helps refine HACID-DSS functionalities, align system design with user needs, and foster trust in its outputs.

- **Primary Metric:** Number of participatory interventions implemented throughout the HACID-DSS development pipeline.
- **Measurement Criteria:**
    - Count of completed participatory interventions, including workshops, stakeholder engagements, and design collaborations.
- **Target:** ≥ 5 participatory interventions, demonstrating consistent stakeholder involvement across different stages of system development.

Data Sources

Participatory activities are documented through multiple sources, including:

- Meeting notes, workshop summaries, and design blueprints.
- Policy recommendations and prototype evaluation results.
- Summaries of stakeholder engagement activities.

Completed and planned interventions include:

- **Deliberative workshops** on values elicitation, risk assessment, and prototype evaluation.
- **Cross-domain deployment workshops** to explore HACID-DSS applications.

- **Collaborations with design student cohorts** on usability and cross-domain adaptation.
- **Policy roundtables** to align HACID-DSS development with regulatory considerations.
- **Iterative design reviews** with user feedback integration.

Procedure

Participatory interventions follow a structured process, ensuring stakeholder input is embedded throughout system development:

1. **Collaborative Planning** – Defining aims, research questions, and intended outcomes.
2. **Material Development** – Creating workshop guides, scenario prompts, and evaluation tools.
3. **Participant Recruitment** – Engaging a diverse set of stakeholders, with targeted outreach strategies to include underrepresented groups.
4. **Implementation & Data Collection** – Conducting participatory sessions using qualitative (e.g., group discussions) and quantitative (e.g., structured surveys) approaches.
5. **Analysis & Integration** – Synthesising insights to inform system refinements, policy briefs, and prototype development.
6. **Feedback & Dissemination** – Sharing results with participants via reports, webinars, and blogs to ensure transparency and ongoing engagement.

Each participatory activity is designed with the following critical questions in mind:[81]

- **Who defines success?** Ensures that success criteria are co-defined with participants, avoiding top-down, developer-driven metrics that risk extractive or performative participation.
- **Whose participation matters?** Proactively identifying and engaging relevant stakeholders, with efforts to include voices that might otherwise be marginalised.
- **What is the intent?** Clarifying whether participation aims to improve model accuracy, foster trust, surface ethical concerns, or promote inclusivity, guiding the choice of methods and engagement depth.
- **How are participants valued?** Considering both intrinsic motivations (e.g., contributing to the common good) and extrinsic incentives (e.g., recognition, skill development) to ensure meaningful, non-exploitative involvement.
- **How is the process closed?** Planning for responsible data handling, participant feedback, and the long-term integration of participatory insights into system development and evaluation.

This approach ensures that participatory interventions are not one-off events but integral components of HACID's design process, fostering continuous learning and adaptation.

---

[81] "Participatory AI for Humanitarian Innovation: A Briefing Paper." *Nesta*, https://www.nesta.org.uk/report/participatory-ai-humanitarian-innovation-briefing-paper/. Accessed 26 Feb. 2025.

Rationale for Approach

Participatory AI moves beyond traditional top-down development, ensuring that AI systems are not only technically robust but also socially responsive, ethically grounded, and contextually relevant. Stakeholder involvement across multiple stages helps enhance system legitimacy, foster trust, and align HACID with real-world user needs.[82] [83]

HACID's Hybrid-CI model integrates both human and machine intelligence, making participatory design essential for ensuring that human expertise remains an evolving, active part of system development. As outlined in recent research, participatory methods can help navigate power dynamics, improve the interpretability and fairness of AI models, and mitigate risks associated with bias and unintended consequences.[84] [85] In HACID, participatory approaches play a role in the following key stages:

- **System Design** – Stakeholders co-design functionalities to address real-world needs.
- **Values Elicitation** – Deliberative processes ensure alignment with ethical and societal expectations.
- **Prototype Evaluation** – Iterative testing and feedback loops improve usability and impact.

This methodology aligns with broader responsible AI principles, emphasising inclusivity, transparency, and accountability in AI development.

2.12.3. Limitations

The participatory design approach effectively integrates diverse stakeholder perspectives, enhancing system relevance and ethical awareness. However, there are several key limitations:

- **Recruiting a diverse participant pool remains difficult,** particularly in specialised domains primarily comprising highly educated professionals—this limits demographic representation.
- **Participation is resource-intensive**, making it challenging to retain contributors for extended engagements.
- **Small-group deliberations risk dominance by authoritative voices**, potentially skewing outcomes.

---

[82] Delgado, Fernando, et al. *The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice*. arXiv:2310.00907, arXiv, 2 Oct. 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2310.00907.
[83] "Participatory AI for Humanitarian Innovation: A Briefing Paper." *Nesta*, https://www.nesta.org.uk/report/participatory-ai-humanitarian-innovation-briefing-paper/. Accessed 26 Feb. 2025.
[84] "Participatory AI for Humanitarian Innovation: A Briefing Paper." *Nesta*, https://www.nesta.org.uk/report/participatory-ai-humanitarian-innovation-briefing-paper/. Accessed 26 Feb. 2025
[85] Delgado, Fernando, et al. *The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice*. arXiv:2310.00907, arXiv, 2 Oct. 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2310.00907.

- **The qualitative nature of participatory interventions constrains scalability**, making generalisable statistical analysis difficult.

2.13. KPI 13 Participatory evaluation

How well did we achieve the goals of participation?

2.13.1. Relevant objective:

OBJ7: Participatory AI: deploy a participatory approach to the development of hybrid collective intelligence exploiting human expertise and AI.

The goal of HACID is to make the HACID-DSS deployable in diverse application sectors through a participatory AI.

2.13.2. Measurement Approach

Methodology for Measurement

Status: In progress

Metrics used

This KPI evaluates stakeholder satisfaction with the participatory process, measuring whether participants feel their contributions were valued, well-integrated, and aligned with best practices in Participatory AI design[86].

- **Primary Metric:** Stakeholder satisfaction with the participatory process.
- **Measurement Criteria:** Survey-based assessment of stakeholder perceptions of key principles of participatory AI design.
- **Target:** High satisfaction and/or meaningful engagement reported by ≥ 80% of stakeholders.

Data Sources

Participant engagement and satisfaction data are collected through:

- **Post-session surveys** distributed to all workshop participants (see Appendix 1 for survey overview).
- **Facilitator notes** capturing real-time participant discussions, providing qualitative insights into engagement.

Procedure

The survey is administered at the end of each participatory activity to capture immediate feedback. Facilitators introduce the survey, explain its purpose, and provide instructions. Participants access the survey via a link, with facilitators available for support. Responses are anonymised to encourage honest feedback. Qualitative insights are captured throughout activities by facilitators and note-takers.

[86] "Participatory AI for Humanitarian Innovation: A Briefing Paper." *Nesta*, https://www.nesta.org.uk/report/participatory-ai-humanitarian-innovation-briefing-paper/. Accessed 26 Feb. 2025.

Rationale for Approach

The survey-based approach ensures a repeatable, quantifiable assessment of stakeholder engagement in participatory activities, allowing for structured comparisons over time while remaining accessible to diverse participants. It draws from frameworks in Participatory AI design, particularly Nesta's Participatory AI for Humanitarian Innovation[87], which outlines five core principles for effective stakeholder engagement:

1. **Who defines success?** – Ensuring participatory processes reflect stakeholder priorities, not just developer-driven goals.
2. **Whose participation matters?** – Identifying and engaging diverse, relevant stakeholders.
3. **What is the intent?** – Clarifying whether participation aims to **improve AI models, foster trust, or ensure ethical considerations**.
4. **How are participants valued?** – Recognising contributions meaningfully, both intrinsically and extrinsically.
5. **How is the process closed?** – Ensuring transparency in how participant input is used.

2.13.3. Limitations

The survey-based approach is efficient and easy to administer, allowing for the systematic evaluation of participatory principles across multiple activities. However, challenges **include**:

- **Small sample sizes** in some workshops, limiting generalisability.
- **Survey item length**—some questions are too detailed for mobile users or time-constrained settings.
- **Limited qualitative insights**, as surveys focus on structured responses rather than open-ended feedback.

[87] "Participatory AI for Humanitarian Innovation: A Briefing Paper." *Nesta*, https://www.nesta.org.uk/report/participatory-ai-humanitarian-innovation-briefing-paper/. Accessed 26 Feb. 2025.

3. Discussion & Conclusions

Review of HACID's KPI evaluation framework.

The HACID evaluation framework builds on existing AI evaluation models, particularly the Human-AI Collaboration (HAIC) Framework and the Participatory AI for Humanitarian Innovation Framework.[88] [89] The HAIC framework's focus on structured criteria for assessing human-AI interaction, aligns well with HACID's decision-support effectiveness KPIs. For instance, KPI-5 (Collective Accuracy) and KPI-7 (Collective Efficiency) correspond to the HAIC's Task Allocation dimension, while KPI-6 (Collective effectiveness) aligns with Interaction, which evaluates how AI communicates uncertainty and fosters user trust. However, the HAIC does not sufficiently address stakeholder alignment, governance, or participatory design—areas covered by the Participatory AI framework. This framework's five principles of participatory AI design are reflected in KPI-10 (Process Evaluation) and KPI-12 (Participatory Design), which address the key participatory design questions: *'Who defines the process?'* and *'Whose participation is required?'*, while KPI-11 (Output Evaluation) addresses *'What is the intent behind participation?'*. Neither framework captured HACID's evaluation needs across both technical and participatory perspectives.

Challenges in Domain-Specific Adaptability

A key challenge in KPI development was ensuring adaptability across HACID's two core domains: climate services and healthcare. While both require accurate, interpretable, and trustworthy decision support, their evaluation priorities differ due to three key factors.

First, artificial intelligence in healthcare is often benchmarked against established ground truths, such as diagnostic accuracy, whereas climate decision support operates in a context of uncertainty and policy-driven priorities. Second, the role of the system varies: in healthcare, it focuses on high-stakes diagnostic reasoning, while in climate services, it integrates scientific predictions with adaptation strategies, where "correct" answers may be unknown. In the latter case, performance evaluation can be further complicated by a lack of clear ground truths and the fact that the effects of recommended actions may take years to materialise in complex environments.

Third, key evaluation priorities emerged during collaborative KPI development with implementing organisations. For instance, the Met Office prioritised alignment with policy frameworks and broader consensus within the scientific literature, whereas Human Dx focused on diagnostic accuracy and expert decision-making. This distinction directly influenced KPI structuring, with decision-support effectiveness KPIs (e.g., KPI 5, KPI 7, KPI 8, KPI 9) being primarily evaluated in the medical domain.

---

[88] Fragiadakis, George, et al. *Evaluating Human-AI Collaboration: A Review and Methodological Framework*. arXiv:2407.19098, arXiv, 9 July 2024. *arXiv.org*, https://doi.org/10.48550/arXiv.2407.19098.

[89] Berditchevskaia, A., Peach, K., and Malliaraki, E. (2021). Participatory AI for humanitarian innovation: a briefing paper. London: Nesta.

Recommendations & Future Directions

HACID's evaluation framework effectively integrates technical, decision-support, and participatory dimensions, ensuring a comprehensive assessment of Hybrid-CI. However, further development is needed to enhance scalability, cross-domain applicability, and the refinement of evaluation metrics—offering opportunities to strengthen HACID's assessment framework in future deployments.

1. **Scalability of participatory evaluations** – While deliberative workshops provide rich qualitative insights, broadening the use of validated digital methods and structured survey-based metrics could enhance participation while reducing resource constraints. Developing scalable participatory approaches will be essential for applying HACID across diverse contexts.
2. **Expanding evaluation across domains –** HACID's decision-support KPIs were primarily validated in the medical domain. Extending similar metrics to climate services and beyond will ensure broader applicability. Future efforts should also focus on devising evaluation strategies for use cases where direct validation opportunities are limited, such as emerging domains with high uncertainty or limited ground truth data.
3. **Refining the transition from values to evaluation metrics** – HACID's framework makes important progress in surfacing and integrating stakeholder values and perceived risks, but further refinement is needed to translate these insights into structured, quantifiable assessment criteria for continuous system evaluation. Developing standardised metrics for values-based evaluation will improve consistency across domains.
4. **Creating benchmarking datasets and repeatable evaluation protocols** – Establishing standardised evaluation protocols and guidelines for development of benchmarking datasets would enable more repeatable and generalisable assessments of HACID's effectiveness, while building an evidence base for its effectiveness over time.[90] [91] This would allow for comparability across use cases and over time, supporting systematic improvements and ensuring robustness in decision-support outcomes.
5. **Real-world validation** – Moving beyond case-based assessments to test HACID in live operational decision-making settings is critical for understanding its practical impact and usability. This includes assessing how HACID functions in high-stakes, time-sensitive environments, as well as its long-term effects on decision-making behaviour, trust, and adoption.

[90] Karargyris, Alexandros, et al. "Federated Benchmarking of Medical Artificial Intelligence with MedPerf." *Nature Machine Intelligence*, vol. 5, no. 7, July 2023, pp. 799–810. *www.nature.com*, https://doi.org/10.1038/s42256-023-00652-2.
[91] Sourlos, Nikos, et al. "Recommendations for the Creation of Benchmark Datasets for Reproducible Artificial Intelligence in Radiology." *Insights into Imaging*, vol. 15, no. 1, Oct. 2024, p. 248. *BioMed Central*, https://doi.org/10.1186/s13244-024-01833-2.

Glossary of Terms

A

- **AI (Artificial Intelligence)** – The simulation of human intelligence by machines, including learning, reasoning, and decision-making capabilities.
- **Annotation** – The process of labelling data (e.g., medical texts, climate datasets) for AI training and evaluation.

C

- **Collective Intelligence (CI)** – The enhanced problem-solving ability that emerges when multiple experts or AI systems collaborate.
- **Competency Questions (CQs)** – Structured questions used to assess the coverage and effectiveness of a Knowledge Graph.

D

- **Decision-Support System (DSS)** – A computer-based tool designed to assist human decision-making by aggregating and analyzing information.
- **Decision Similarity –** The extent to which an individual's suggested solutions align with those of others in a group, often used to assess collective decision-making dynamics.
- **Deliberative Workshop** – A structured discussion format where stakeholders explore values, risks, or system designs in AI development.
- **Domain Expertise** – Knowledge and experience specific to a particular professional or scientific field.

E

- **Explainability** – The ability of an AI system to provide understandable reasons for its outputs and decisions.
- **Expert Aggregation** – The process of combining multiple expert opinions to improve decision-making accuracy.

H

- **HACID (Hybrid Collective Intelligence for Decision-making)** – A research initiative integrating human expertise and AI to improve decision-making in high-stakes domains.
- **HAIC (Human-AI Collaboration) Framework** – A structured evaluation framework assessing AI-human interaction in decision-making.
- **Human Dx (Human Diagnosis Project)** – A collaborative platform for medical professionals to generate and validate diagnoses.

K

- **Key Performance Indicator (KPI)** – A measurable value used to track progress toward specific project objectives.
- **Knowledge Engineering (KE)** – The process of structuring, organizing, and integrating knowledge into AI-driven decision-support systems.
- **Knowledge Graph (KG)** – A structured database representing relationships between entities to improve AI reasoning and retrieval.

L

- **Large Language Model (LLM)** – A type of AI trained on vast datasets to understand and generate human-like text (e.g., GPT-4, Claude, Gemini).

M

- **Metadata Boosts** – The enhancement of decision accuracy by incorporating additional expert-provided data, such as response time or confidence levels.
- **MIMIC-IV (Medical Information Mart for Intensive Care)** – A large, publicly available medical dataset used in AI research.

N

- **Nesta's Participatory AI Framework** – A set of guidelines ensuring AI systems are co-developed with stakeholder input.
- **Natural Language Processing (NLP)** – A field of AI that enables computers to understand, interpret, and generate human language.

O

- **Ontology** – A structured representation of knowledge within a domain, often used in AI for reasoning and information retrieval.
- **OWLUnit** – A tool for evaluating Knowledge Graphs through competency question verification.

P

- **Participatory AI (PAI)** – An approach that ensures AI development includes meaningful input from diverse stakeholders.
- **Process Evaluation** – A method for assessing whether AI evaluation criteria align with stakeholder needs and values.

R

- **Retrieval-Augmented Generation (RAG)** – An AI method that combines document retrieval with language model generation to improve accuracy and reduce hallucinations.
- **Risk Assessment Workshop** – A participatory activity where stakeholders identify and evaluate risks associated with AI decision support.

S

- **SNOMED Clinical Terms (SNOMED-CT)** – A comprehensive, standardized medical terminology system used for AI-driven medical diagnosis.
- **SPARQL** – A query language for retrieving and manipulating data stored in a Knowledge Graph.

T

- **Task Allocation** – The division of responsibilities between AI and human decision-makers in hybrid systems.
- **Technical Robustness** – The reliability and security of an AI system, ensuring it functions as expected under various conditions.
- **Trustworthiness** – The extent to which AI systems are perceived as reliable, transparent, and aligned with ethical standards.

U

- **User Engagement Scale (UES)** – A validated tool for measuring user interaction and satisfaction with AI-based systems.
- **Uncertainty Awareness** – AI's ability to communicate the degree of uncertainty in its outputs.

Appendix

Appendix 1: Participatory AI Survey Items

This appendix presents the survey items used to assess stakeholder satisfaction and engagement in participatory activities within the HACID project. The survey was designed based on Nesta's Participatory AI for Humanitarian Innovation framework, ensuring alignment with established principles of inclusive, transparent, and meaningful participation.

The survey items were administered at the end of participatory activities and structured around five key dimensions of effective participatory AI design:

Survey Items

Participants were asked to rate the following statements on a **5-point Likert scale** (Strongly Disagree – Strongly Agree):

1. **Perceived Influence on Outcomes:**
    - "Values, opinions, and ideas were considered during this process."
2. **Inclusivity and Diversity:**
    - "The selection of participants was appropriate for the objectives."
3. **Clarity and Purpose:**
    - "The goals of our participation were clear and aligned with my understanding."
    - "The activities provided an opportunity to make a meaningful contribution."
4. **Recognition and Appreciation:**
    - "Participation was valued and appropriately recognised."
5. **Transparency in Decision-Making:**
    - "I have a clear understanding of how my contributions will be used."

Survey Implementation

- The survey was administered digitally via a secure link at the end of each workshop or participatory session.
- Responses were anonymised to encourage honest feedback.
- Facilitators were available to provide **clarification** and ensure participants understood each question.

These survey results provided a quantifiable measure of participatory engagement across multiple activities, informing the evaluation of KPI 13 (and also KPIs 10 & 11) and supporting broader assessments of stakeholder involvement in the HACID-DSS development process.