# IRCAI

**Scientific excellence, global public dialogue and technology for sustainability, inclusion and equality.**

# Responsible AIoT and TinyML

## School on Applied AI for Sustainable Development | (smr 4210)

Trieste 24.3.2025

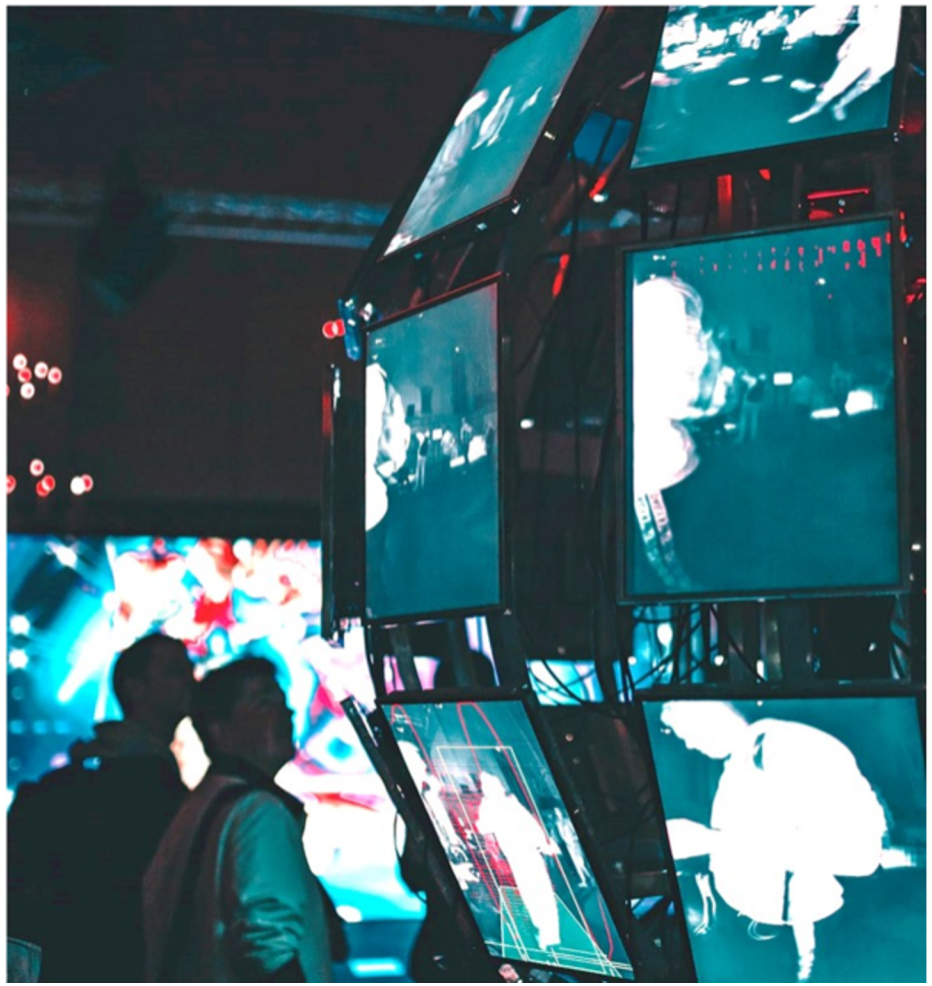
unesco
Centre
Under the auspices
of UNESCO

ICTP The Abdus Salam International Centre for Theoretical Physics

IRC AI
International Research Centre
of Artificial Inteligence
under the auspices of UNESCO

RAIDO

Co-funded by the European Union



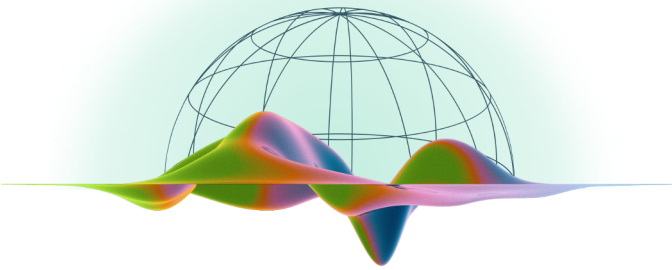
Potential for new and old threats being enhanced

- Citizen control as envisaged in Brave New World
- Automatic processing of decisions that ignore marginalised groups
- Automatic weapons that can be used to disable legitimate protest or wage destructive wars


Recommendation on the Ethics of AI

- After extensive consultations, recommendation adopted at the UNESCO assembly in November 2021
- Human centered: how can individual humans be protected from misuse and enriched by the use of AI
- Closely aligned with European humanistic values and with EU initiatives as well as Council of Europe

**Global Forum on the Ethics of AI (GFEAI)**



unesco Centre Under the auspices of UNESCO

IRC AI International Research Centre of Artificial Intelligence under the auspices of UNESCO





unesco
Centre
Under the auspices
of UNESCO

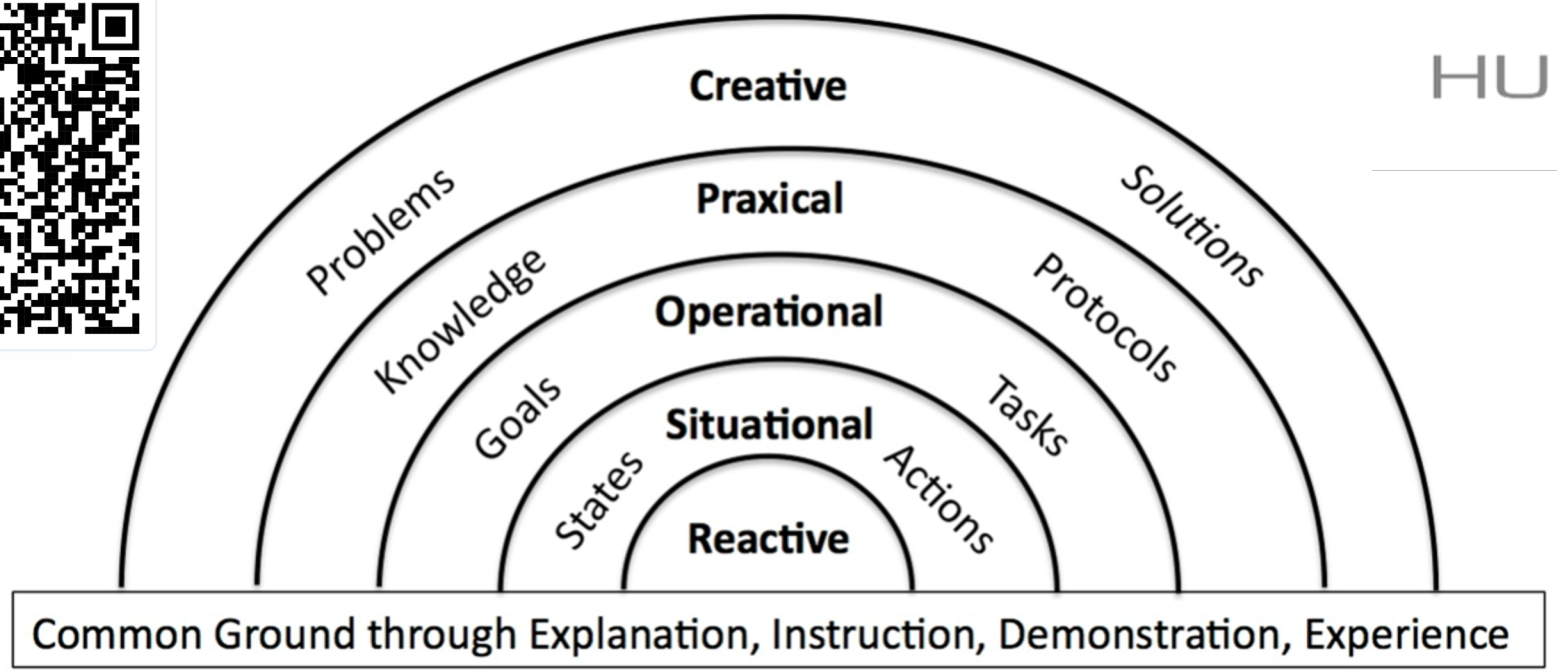
IRC AI
International Research Centre
on Artificial Intelligence
under the auspices of UNESCO

- EU AI Act is first attempt, but risk based so that many systems not covered
- One advantage of the UNESCO recommendation is that it is vague: forcing designers to think through unforeseen consequences:
  - assessing when human rights have been compromised, system acted unfairly, introduced bias, etc.
  - Any formal definition will inevitably be limited and motivate circumvention

unesco
Centre
Under the auspices
of UNESCO

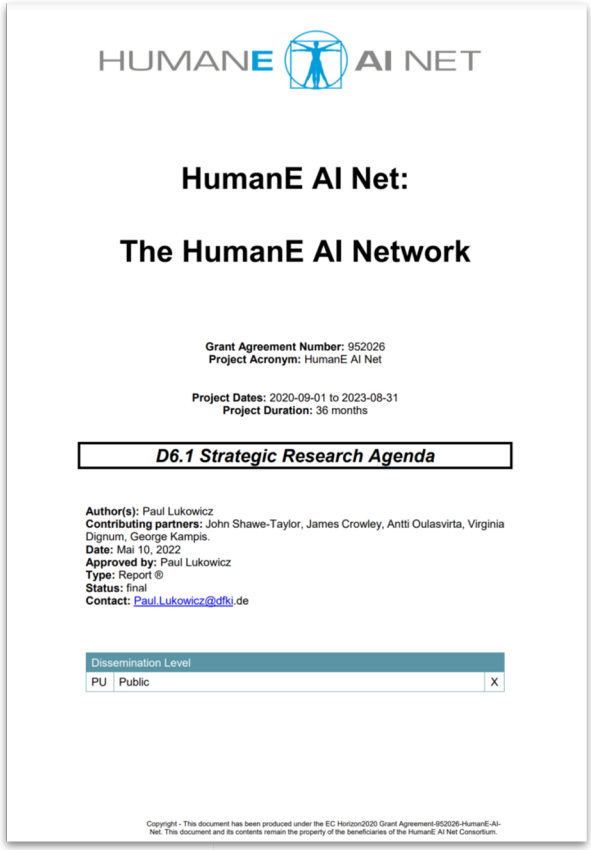
IRC AI
International Research Centre
on Artificial Intelligence
under the auspices of UNESCO


- Human-centric AI is at the centre of the European vision of a positive role for AI
- AI that empowers humans to be more effective, more creative, more understanding
- Human-centric AI (HCAI) has been the focus of the Humane AI Network of Excellence
- How does its research agenda envisage HCAI?




Humane AI Research agenda highlights the ingredients of Collaborative Intelligent Systems:

- Need to find ‘common ground’ across a range of levels in order to enable effective cooperation/communication
- Levels identified roughly correspond to different styles of collaboration with collaborative systems potentially involving more than one level



HUMANE AI NET

**HumanE AI Net:**

**The HumanE AI Network**

**Grant Agreement Number:** 952026
**Project Acronym:** HumanE AI Net

**Project Dates:** 2020-09-01 to 2023-08-31
**Project Duration:** 36 months

***D6.1 Strategic Research Agenda***

**Author(s):** Paul Lukowicz
**Contributing partners:** John Shawe-Taylor, James Crowley, Antti Oulasvirta, Virginia Dignum, George Kampis.
**Date:** Mai 10, 2022
**Approved by:** Paul Lukowicz
**Type:** Report ®
**Status:** final
**Contact:** Paul.Lukowicz@dfki.de

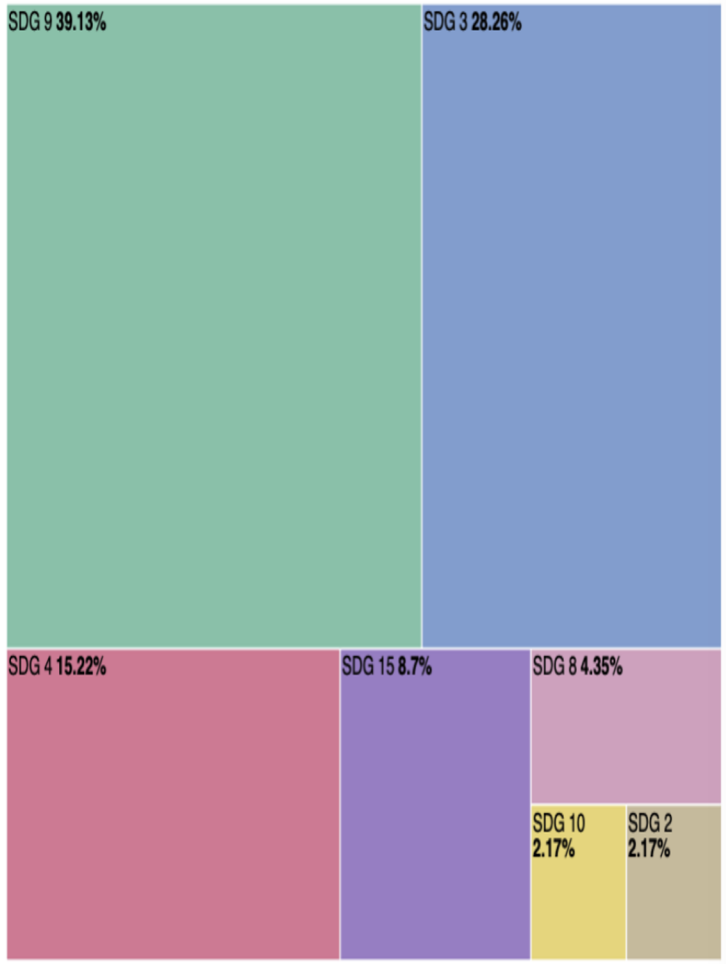
| Dissemination Level | | |
|---|---|---|
| PU | Public | X |



unesco Centre Under the auspices of UNESCO

IRC AI International Research Centre of Artificial Intelligence under the auspices of UNESCO

# Responsible AI in TinyML and STEAM Education

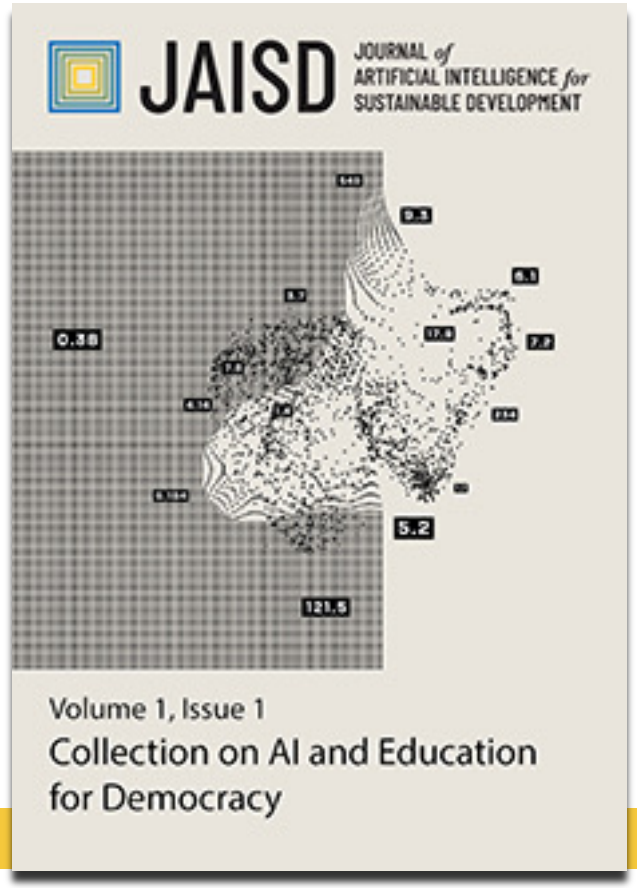



OPERATIONALIZED ETHICAL AND FUNCTIONAL DIMENSIONS WITH PRIORITY AREAS FOR TINYML AND STEAM EDUCATION

| Dimension | Description | EdgeAI/TinyML | STEAM Education |
|---|---|---|---|
| **Transparency** | Clearly communicate AI operations, data handling, and inference logic to participants to build trust and accountability in both community science and educational contexts. | High – enables interpretability of local models and data flows. | High – critical for teaching explainability and responsible AI use. |
| **Inclusivity** | Ensure equitable access to AI tools, datasets, and learning resources, addressing disparities in infrastructure, gender representation, and cultural participation. | Medium – supports broader participation in device deployment. | High – foundational for equitable AI literacy and engagement. |
| **Fairness** | Identify and mitigate algorithmic bias across data collection, labeling, and model evaluation processes, especially in community or low-resource settings. | High – essential for localized and context-aware model training. | High – core concept for teaching ethical reasoning in AI and data science. |
| **Environmental Responsibility** | Assess and reduce lifecycle impacts of AI hardware, including sourcing, energy consumption, and end-of-life management. | High – TinyML's edge devices directly affect ecological sustainability. | Medium – integrated into STEAM through sustainability modules and project design. |
| **Cultural Diversity** | Incorporate local knowledge, indigenous epistemologies, and linguistic diversity into AI datasets, interpretations, and learning activities. | Medium – supports culturally relevant sensor deployment and data interpretation. | High – encourages interdisciplinary, culturally responsive pedagogy. |
| **Operational Assistance** | TinyML devices perform localized inference for environmental monitoring, health tracking, or IoT sensing, reducing reliance on cloud infrastructure. | High – primary operational focus of edge ML systems. | Medium – serves as an applied learning example in engineering and computing courses. |
| **Decision Support** | Generative AI systems synthesize data insights and summaries to assist collective decision-making and policy formation in citizen science projects. | Medium – complements TinyML data with higher-level synthesis. | High – enhances inquiry-based learning, problem-solving, and critical reflection. |
| **Interpretive / Narrative Collaboration** | AI systems co-generate visualizations, reports, and educational materials with human collaborators, promoting reflective learning and knowledge co-creation. | Medium – relevant in projects using embedded visualization or reporting tools. | High – central to creative and narrative integration in STEAM curricula. |


One way to try to offset the potential misuse is to regulate greater disclosure

- AI should be educational, broadening our understanding of situations and content
- This could be a requirement that enables users to make up their own minds
- Many of the manipulations rely on suggesting that full disclosure is not available


**Informed Consent:** ensure participants know why their data is being collected, how it will be used, and any associated risks.

**Privacy**: Protect the identities of participants by anonymizing or de-identifying data.

**Transparency**: clear purpose of the data collection, the methods used, and potential benefits and risks.

**Data Sharing**: participants should be aware and have given consent.

**Vulnerable Populations**: extra precautions when collecting data from children, the elderly, and other vulnerable groups.

**Potential Harm**: Assess and minimize potential risks to participants. This includes emotional distress, financial harm, or other adverse effects.

**Purpose Change**: seek fresh consent from participants.

**Limitation**: avoid over-collection.

**Accuracy**: data is accurate and represents the truth.

**Beneficence**: data collection should be for good.

**Models vs. Data:** storing models, forgetting the data

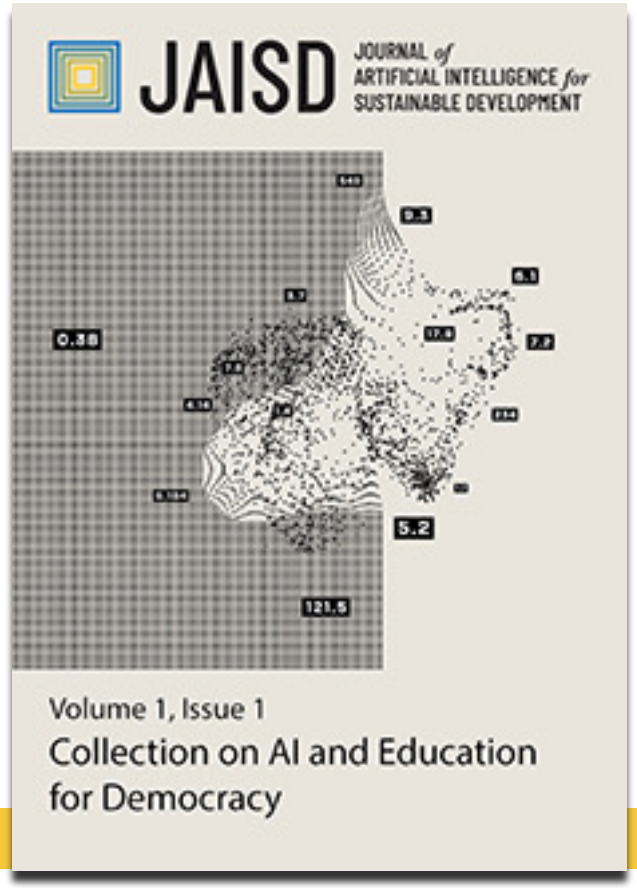
**Access**: who has access to the data.

**Long-term Storage**: ensure secure storage methods.

**Feedback**: offer participants feedback or results from the study if they express interest.

**Regulation and Legislation**: adhere to local, national, and international data protection laws and regulations.

**Cultural Sensitivity**: understand and respect cultural norms and traditions when collecting data


**Transparency**: AI algorithms should be transparent in terms of how they work and how decisions are made – XAI.

**Bias and Fairness:** It's crucial to identify, reduce, and disclose biases to ensure fairness in decision-making.

**Accountability**: determine who is responsible for AI decisions.

**Privacy**: protect the privacy of individuals when processing their data.

**Data Security**: safeguard this data against breaches and unauthorized access.

**Autonomy**: People should always have a choice, especially concerning decisions that significantly affect their lives.

**Informed Consent**: Users should be informed when their data is being processed by AI and have a clear understanding of how the AI system uses their data.

**Explainability**: AI decisions should be interpretable and explainable.

**Generalizability and Robustness**: AI models are not overfitted to their training data and can generalize well to real-world situations.

**Economic and Social Impacts**: Consider the broader societal and economic effects of AI for good

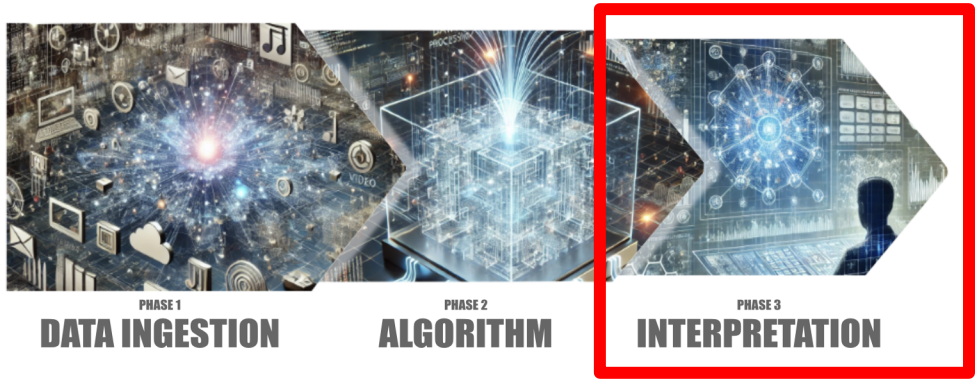
**Long-term Considerations**: longer-term implications of AI, including the potential for systems to evolve or be used in ways not originally intended.

**Environmental Impact**: consider the environmental footprint and strive for sustainable AI research.

**Continual Monitoring**: AI systems, especially those deployed in dynamic environments, should be continuously monitored to ensure they are behaving as expected.

**Stakeholder Participation**: Involve relevant stakeholders, including those affected by AI systems, in the design, development, and deployment processes.

**Regulation and Standards:** Align AI practices with existing regulations and contribute to the development of ethical standards for AI.





**Avoiding Bias**: confirmation bias (favoring information that aligns with one's existing beliefs) or availability bias (relying on immediate examples).

**Transparency**: clearly communicate the methods and tools used in the analysis so that others can understand and potentially reproduce the results.

**Overreach**: Do not make claims or conclusions that go beyond what the data actually supports.

**Accuracy**: Ensure that interpretations are based on accurate and robust findings. Avoid cherry-picking data or results to support a particular narrative.

**Full Disclosure**: Present any limitations, uncertainties, or assumptions associated with the analysis. This allows for a more nuanced understanding of the results.

**Conflict of Interest**: Be transparent about any potential conflicts of interest that might influence the interpretation or presentation of the results.

**Sensitivity**: Be aware of the broader social and cultural context when interpreting data.

**Feedback Loop**: Consider the potential feedback loop effects where the interpretation of data analysis can influence future data collection and results.

**Stakeholder Consideration**: Think about who is affected by the interpretation of the data and ensure that their interests and perspectives are considered.

**Cultural Context**: Especially in global or multicultural studies, understand the cultural nuances and be wary of interpreting data through only one cultural lens.

**Ethical Implications**: Consider the broader ethical implications of any conclusions drawn, especially if they may lead to actions or policies that impact people's lives.

**Peer Review**: Encourage peer review of findings and interpretations to ensure validity and mitigate potential biases.

**Translating to Policy**: If data interpretation is used to inform policy or business decisions, ensure that the recommendations are ethically sound, justifiable, and in the best interest of those affected.

**Honesty**: Always approach data interpretation with honesty, and avoid the temptation to manipulate or selectively present results for personal or organizational gain.

**Continual Learning**: Recognize that interpretations might change as more data becomes available or as methods evolve. Be open to updating or revising interpretations in the light of new evidence.

Continuous Compliance with Legislations for Privacy, IRP & Data Collection

AI Ethics & Princples by Design Framework

Reinforced Benchmarking & Feedback-based Progress Monitoring

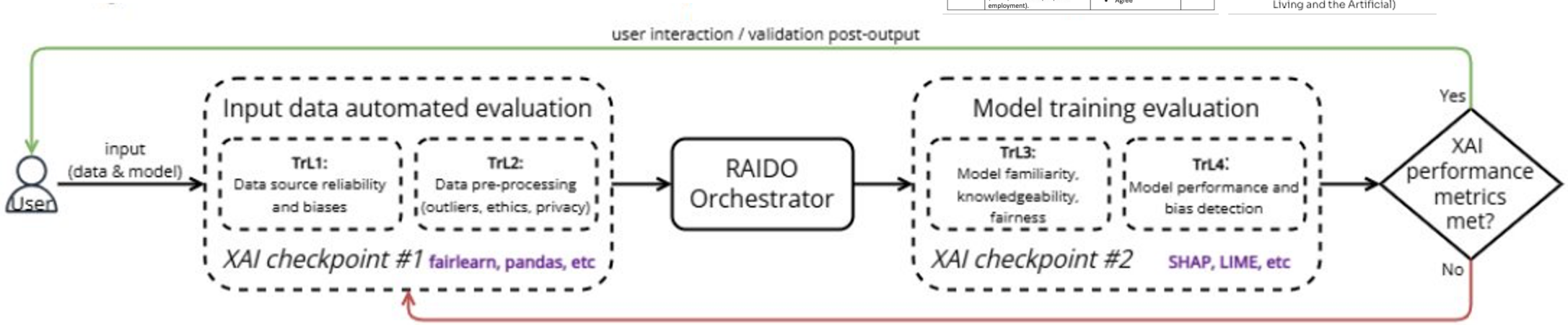
| Acceptance of AI use in life and job/industry | | | |
|---|---|---|---|
| 20 | I am interested in exploring new technological developments for my field of work/in my daily life. | • Strongly Disagree<br>• Disagree<br>• Neutral/Neither Agree nor Disagree<br>• Agree<br>• Strongly Agree | |
| 21 | I could imagine including AI in my current work or in my daily life. | • Strongly Disagree<br>• Disagree<br>• Neutral/Neither Agree nor Disagree<br>• Agree<br>• Strongly Agree | |
| 22 | I see the benefit of using AI tools in my current work or in my daily life. | • Strongly Disagree<br>• Disagree<br>• Neutral/Neither Agree nor Disagree<br>• Agree<br>• Strongly Agree | |
| 23 | Policymakers strongly support AI. | • Strongly Disagree<br>• Disagree<br>• Neutral/Neither Agree nor Disagree<br>• Agree<br>• Strongly Agree | |
| 24 | **(For respondents in employment)** The use of AI in my field aligns with what society considers appropriate for the industry. | • Strongly Disagree<br>• Disagree<br>• Neutral/Neither Agree nor Disagree<br>• Agree<br>• Strongly Agree | |
| 25 | **(For respondents in employment)** I am the person who decides on the use of AI at my job. (For people in employment) | • Strongly Disagree<br>• Disagree<br>• Neutral/Neither Agree nor Disagree<br>• Agree<br>• Strongly Agree | |
| 26 | **(For respondents in employment)** I feel that my organisation would support me in the adoption of AI. (New question for people in employment). | • Strongly Disagree<br>• Disagree<br>• Neutral/Neither Agree nor Disagree<br>• Agree | |

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

vito

**VANESSA NUROCK**

Professor of Philosophy at the Université de Côte d'Azur (France) and UNESCO EVA Chair (Ethique du Vivant et de l'Artificiel /Ethics of the Living and the Artificial)



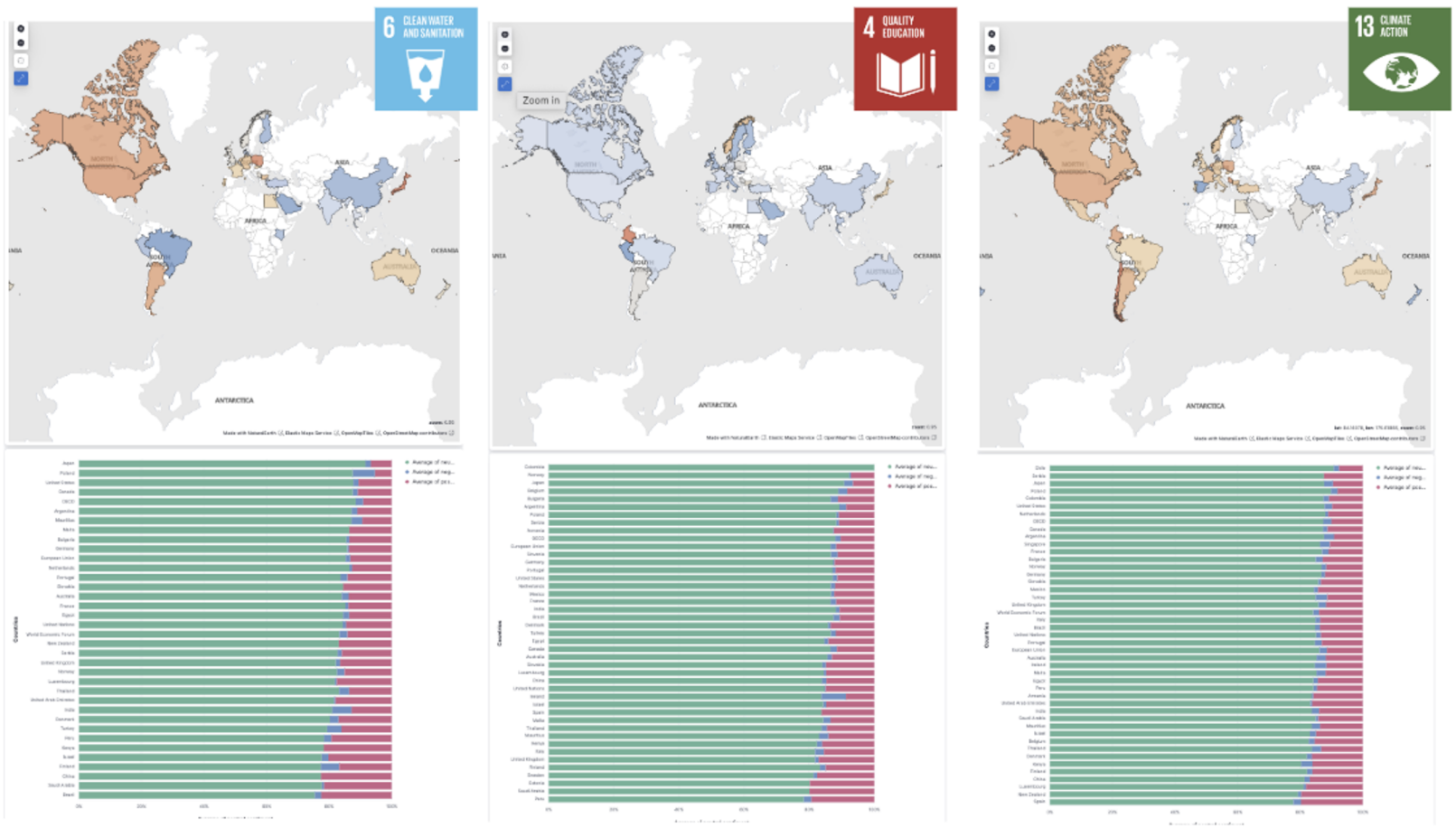

Framework for AI Explainability, Transparency & Trustworthiness

unesco Centre Under the auspices of UNESCO

IRC AI International Research Centre of Artificial Intelligence under the auspices of UNESCO

- ➔ Compute sentiment analysis (using VADER) on title + text of OECD AI policies translated
- ➔ Compare geographically the neutral sentiment on the AI policies to identify tendencies
- ➔ Differentiate between SDGs to explore the variability across the 17 targets and their topics

**Sentiment Bias Aware.**




AI4Gov
Trusted AI for Transparent Public Governance fostering Democratic Values

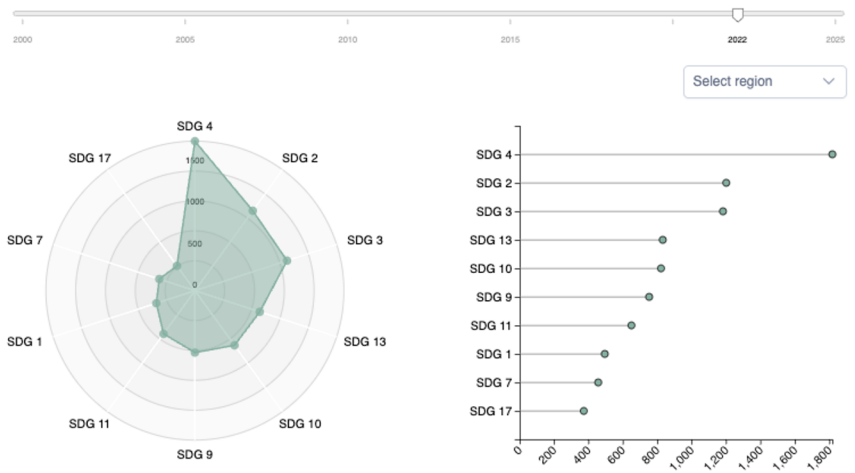
ai4gov-project.eu 101094905

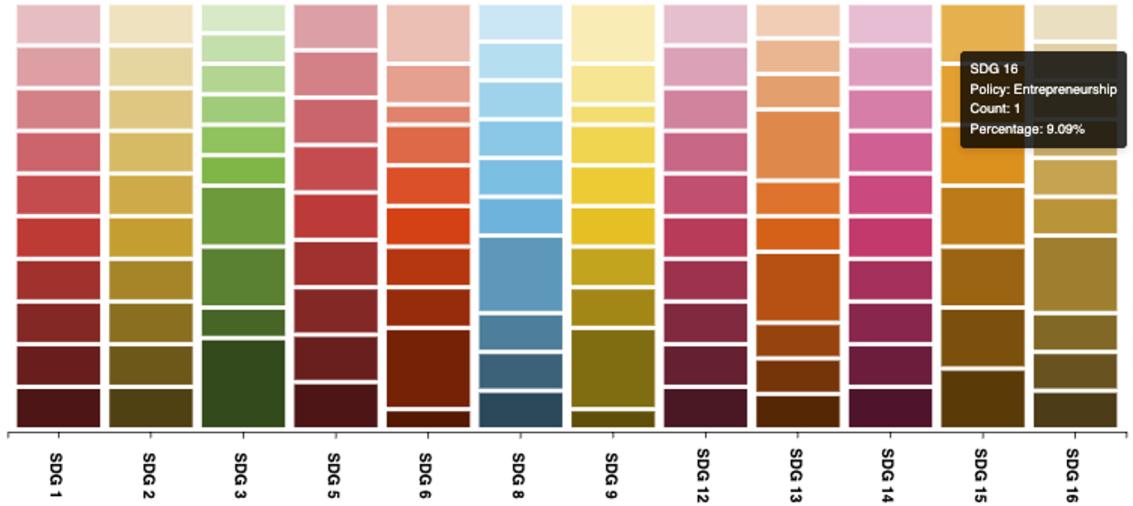
unesco Centre Under the auspices of UNESCO

IRC AI International Research Centre of Artificial Intelligence under the auspices of UNESCO

- ➔ Evaluate the amount of SDG-classified topics across time
- ➔ Compare by geographic world regions to explore prioritisation allowing for comparison
- ➔ Encode the SDG topics based on wikidata concepts identified in the textual documents independently of their language

**Coverage Bias Aware.**




RAID

raido-project.eu

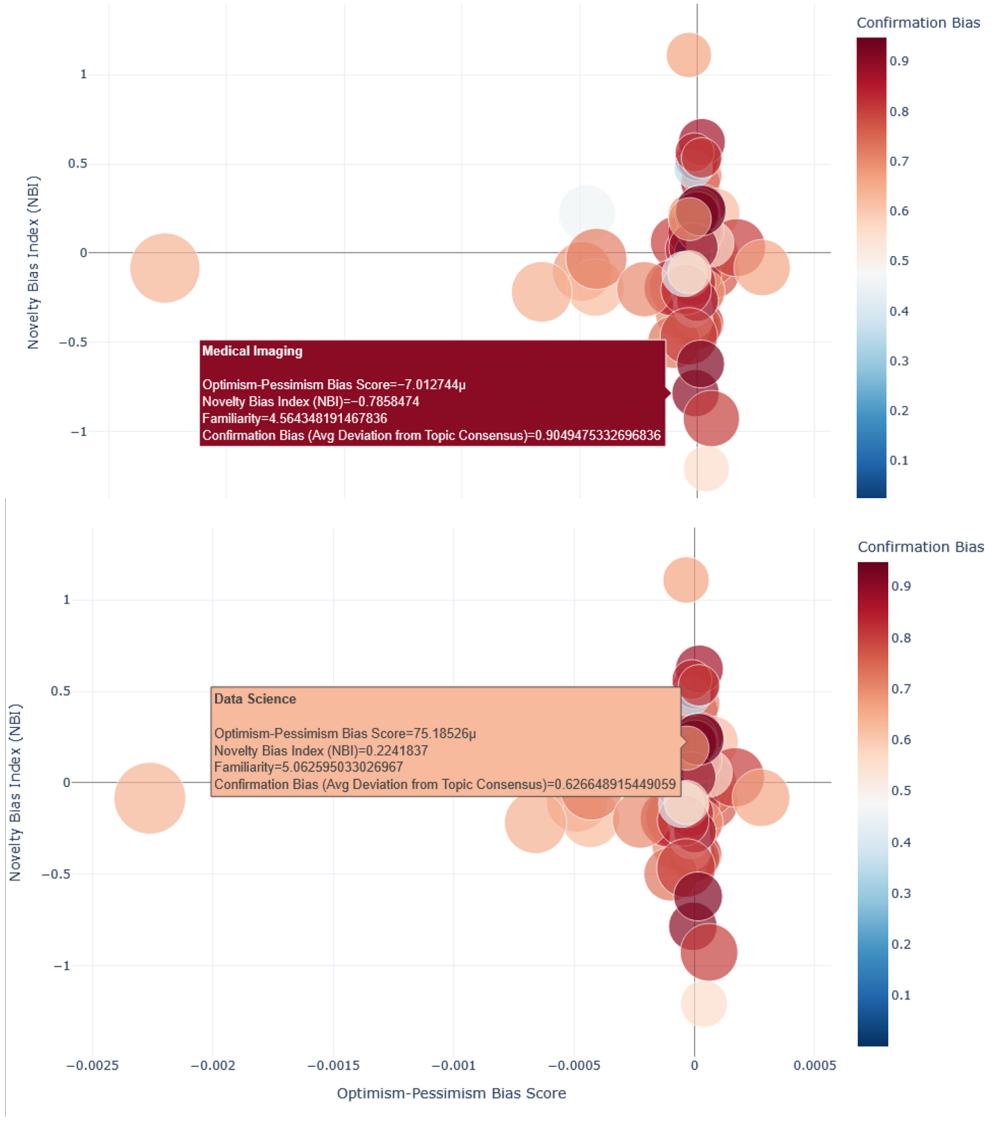
101135800

unesco Centre Under the auspices of UNESCO

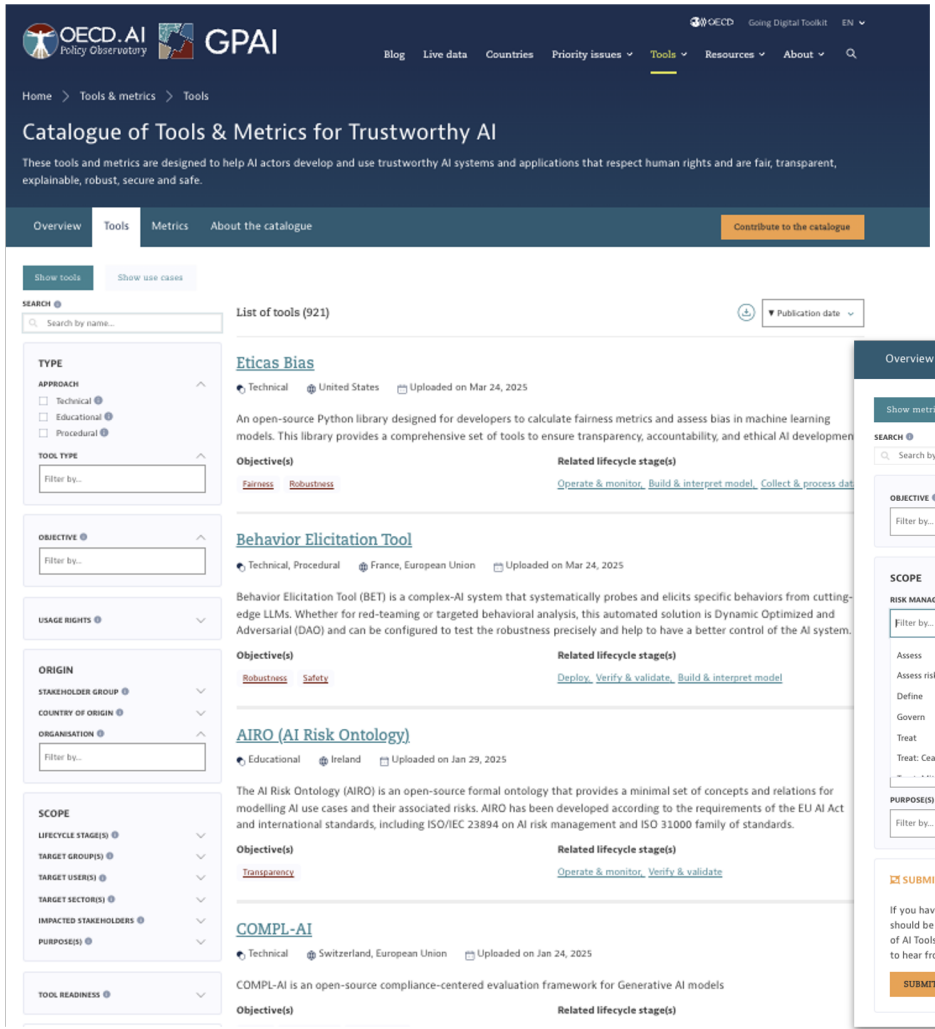
IRC AI International Research Centre of Artificial Intelligence under the auspices of UNESCO

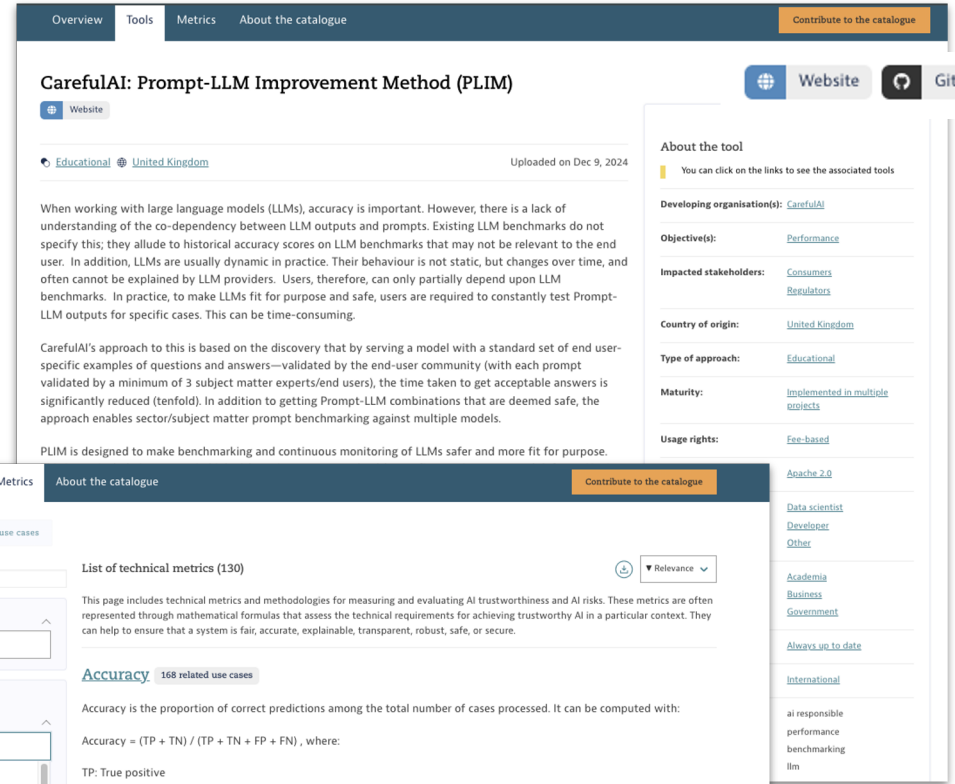
- Integration of cognitive science with computational social science.
- Empirical framework for bias detection in AI discourse.
- Use of semantic linking to align media content with AI taxonomy.
- Identification of cognitive mechanisms shaping AI adoption narratives.
- First large-scale framework connecting cognitive biases to AI innovation adoption.

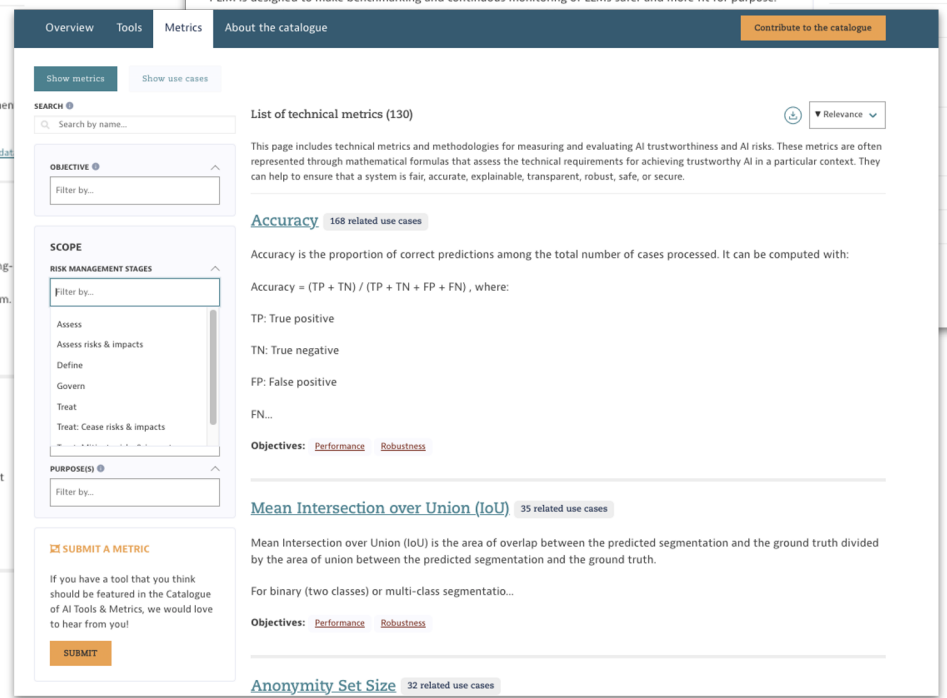
**Cognitive Bias Aware.**



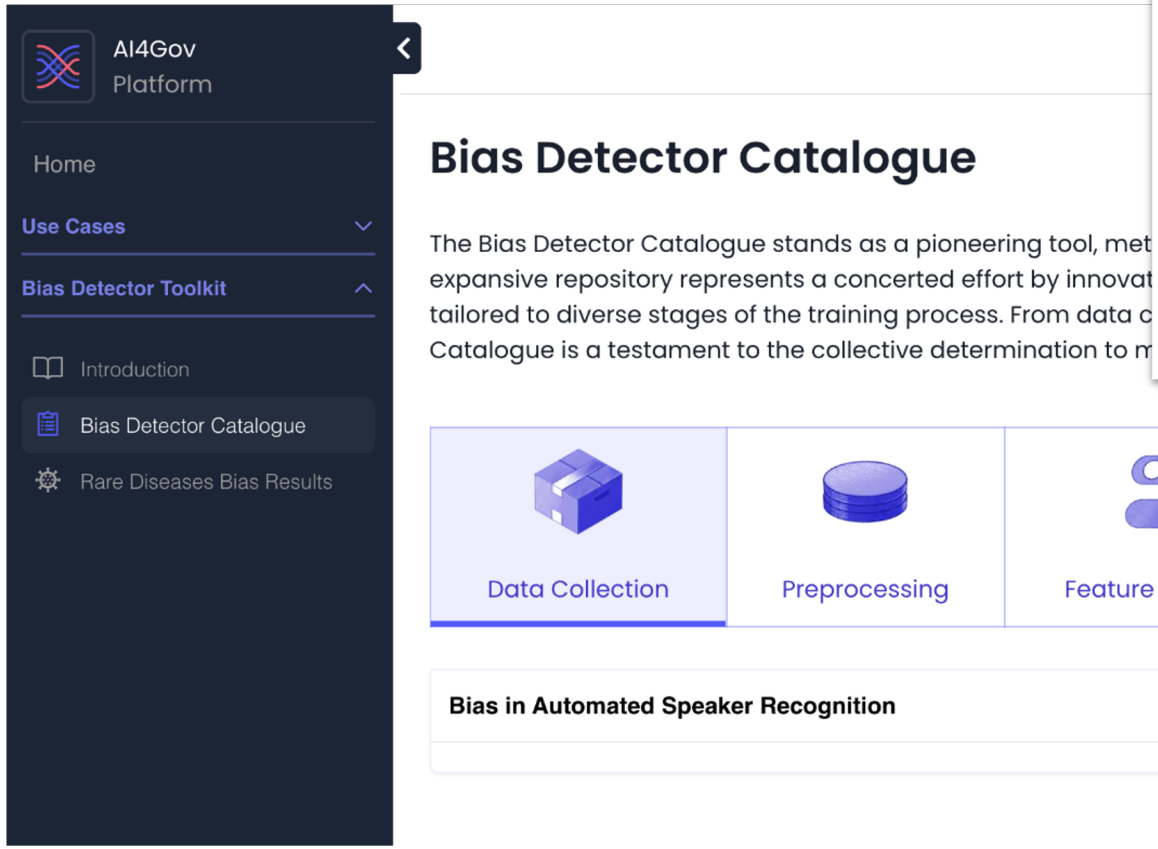

elias-ai.eu 101120237




https://oecd.ai/en/catalogue/tools




# Bias Detector Catalogue

The Bias Detector Catalogue stands as a pioneering tool, met
expansive repository represents a concerted effort by innovat
tailored to diverse stages of the training process. From data c
Catalogue is a testament to the collective determination to m


**Bias in Automated Speaker Recognition** Accuracy: MODERATE Cost: MODERATE

**AIF360: AI Fairness 360 toolkit** Accuracy: HIGH Cost: LOW

**FairMLHealth** Accuracy: UNKNOWN Cost: LOW

**Source:** https://github.com/KenSciResearch/fairMLHealth **Type:** MITIGATION **Programming Language:** PYTHON

**Description:**
FairMLHealth is a healthcare-specific tool for bias analysis. It provides machine-learning fairness, healthcare applications, and variation analysis.

**Applicability:** HEALTHCARE

**Limitations:**
The 'fair' range to be used for these metrics requires judgement on the part of the analyst.

**References:**
Ahmad et al., (2020). Fairness in Machine Learning for Healthcare. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. https://doi.org/10.1145/3394486.3406461.

**Mitigating Unwanted Biases with Adversarial Learning** Accuracy: HIGH Cost: LOW

**Themis-ml: A Fairness-Aware Machine Learning Interface for End-To-End Discrimination Discovery and Mitigation** Accuracy: UNKNOWN Cost: UNKNOWN

**Bias in Automated Speaker Recognition** Accuracy: MODERATE Cost: MODERATE

**Bias Assessment Metrics and Measures** Accuracy: UNKNOWN Cost: UNKNOWN

**Biaslyze** Accuracy: UNKNOWN Cost: LOW

AI4Gov
Trusted AI for Transparent Public Governance fostering Democratic Values

https://cluster-ai4gov.euprojects.net/

unesco Centre Under the auspices of UNESCO

IRC AI International Research Centre of Artificial Intelligence under the auspices of UNESCO

unesco
Centre
Under the auspices
of UNESCO

IRC AI
International Research Centre
of Artificial Inteligence
under the auspices of UNESCO

**International Research Centre on Artificial Intelligence under the auspices of UNESCO (Category II)**

Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana
E: info@ircai.org | W: https://ircai.org/

Joao Pita Costa
in/joaopitacosta/