BUILDING RESPONSIBLE AI ECOSYSTEMS: FROM THEORY TO ACTION FOR PUBLIC GOOD

GITEX ASIA

Singapore, 9.4.2026

ELIAS
unesco
Centre
Under the auspices
of UNESCO
NAiXUS
IRC AI
International Research Centre
of Artificial Inteligence
under the auspices of UNESCO
AI
IN AFRICA
GITEX
ASIA
Singapore
AI
EVERYTHING
–SINGAPORE–

IRCAI: Who are we and what do we do?

IRCAI VIRTUAL LAUNCH

https://ircai.org/launch-2021/

- LAUNCH publicly in March 29 and 30 2021 with **1083 registered participants** from **123 countries** as an ecosystem of research institutions, companies, start-ups and non-profit players.
- Partially founded by the Government of the Republic of Slovenia and has its headquarters at the **Jožef Stefan Institute** in **Ljubljana, Slovenia**.
- Connections to **160 private companies**, **76 public entities** (Foundation/ non-profit), **21 public (University)**; **6 public (Government)**, and **6 Public (NGO),** helped/ing preparing applications for IRCAIs in other regions **(Morocco, Saudi Arabia, Ghana, New Zealand)**
- IRCAI's director **John Shawe-Taylor** is Chairholder of the **UNESCO Chair in Artificial Intelligence** within the UCL doing work in AI and acts as the trustee of the Knowledge 4 All foundation doing work in the Global South.

INTERNATIONAL RECEARCH CENTRE FOR AI
UNDER THE AUSPICES OF UNESCO

Policies Innovation
(UNESCO, Council of Europe, OECD, EU, GPAI, National policies, D4D)

Research on consequences of AI
(Legal frameworks, Ethics in AI, Added Value Models, Jobs,…)

Projects addressing global challenges with AI (Global Ground Truth, Open Education)

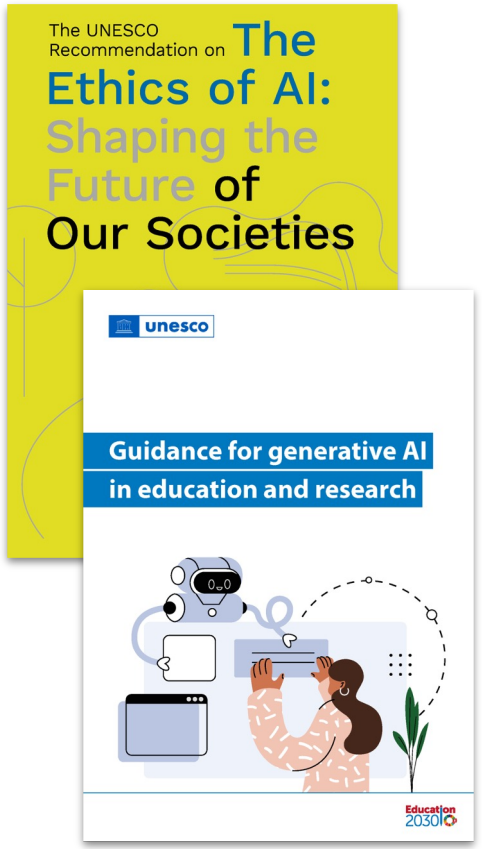
Large Open Analytics Infrastructures
(SDG Observatories, Open Libraries)

Verification programme
(clearing house, mediation verification programme, validation of the ethicality of algorithms and data management)

Financing programme
(Social Impact Bonds, venture capital to finance research and acceleration of start-ups)

Capacity building, Awareness raising
(Journal, AI Olympics, Open education, Events, Formal Educational Programs on AI)

Global Network of AI competences
(NAIXUS, IRCAI backbone network)



ELIAS

unesco Centre Under the auspices of UNESCO

NAiXUS

IRCAI International Research Centre of Artificial Intelligence under the auspices of UNESCO

AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

NAiXUS

naixus.net

NETWORK FOR ARTIFICIAL INTELLIGENCE, KNOWLEDGE AND SUSTAINABLE DEVELOPMENT

FOCUSING ON SCIENTIFIC / TECHNOLOGICAL MAJOR CHALLANGES

Focus on scientific or technological major challenges, with the primary goal to reinforce capacity and progress in critical technologies

SPREADING ADVANCED KNOWLEDGE

Building on existing efforts of already existing platforms, networks and projects, the network will develop mechanisms to spread the latest and most advanced knowledge to all the AI-labs in the United Nations five geographical regions: Africa, Americas, Asia and the Pacific, Europe and Central Asia, and the Middle East and prepare the next generation of talent in AI. Such mechanisms will be defined with all network partners

DEVELOPING SYNERGIES AND CROSS-FERTILIZATION

Another objective is also to develop synergies and cross-fertilization between industry and investors in the network of excellence centres, in particular through internships of academic staff (at all levels) in industry, or PhD programmes with industry;

BECOMING A VIRTUAL LABORATORY

The network will form a common resource and will become a shared facility, as a virtual laboratory offering access to knowledge and expertise and attracting scientists, investors, policymakers and new talents. It should become a reference, creating an easy entry point to AI excellence across the Global South as well and should also be instrumental for its visibility

CREATING A DATA DRIVEN BENCHMARK SYSTEM

Creating a data driven benchmark system for the measurement of SDGs to further create a social impact investment ecosystem for AI research and companies

unesco Centre Under the auspices of UNESCO

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

NTU IoX Center

RAIL LAB

cetic.br Regional Center for Studies on the Development of the Information Society under the auspices of UNESCO

MASAKHANE Community Foundation

AIMS African Institute for Mathematical Sciences SOUTH AFRICA

NORTHEASTERN CIVIC A.I. LAB

DATA-POP ALLIANCE

ellis ALICANTE

DSN Data Scientists Network

ANDORRA RESEARCH + INNOVATION

UNIVERSITY OF GOTHENBURG

K4A knowledge for all

University of Essex

The Alan Turing Institute

TU Delft

QUT Queensland University of Technology

UNIVERSITEIT VAN PRETORIA UNIVERSITY OF PRETORIA YUNIBESITHI YA PRETORIA

ELIAS

unesco Centre Under the auspices of UNESCO

NAiXUS

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

AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

IRCAI AWARD

Adriana-Eufrosina Bora
Project AIMS
(Artificial Intelligence against Modern Slavery)

IRCAI JOURNAL

IRCAI LIBRARY

Find out more at ircai.org

ELIAS

unesco
Centre
Under the auspices of UNESCO

NAiXUS

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

IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

Paid - AI Summit

PANEL DISCUSSION | RESPONSIBLE AI IN ACTION: ACCELERATE AI FOR PUBLIC GOOD IN ASIA

As agentic AI expands across Asia, the key challenge is not scale alone, but relevance, inclusion, and trust. Frugal edge AI and participatory citizen science offer…

5:15PM - 5:55PM

Main Stage

Panel Discussion | Building the AI Vanguard: From Classroom Literacy to Olympiad Excellence

This panel explores the future of AI education across diverse global ecosystems. Moderated by the lead of South Africa's National AI Olympiad, the panel unites…

11:50AM - 12:15PM

10X Stage

Keynote | AI's Global Impact: Reshaping Society and Education

11:40AM - 11:50AM

10X Stage

Senior Showcase & Pitch Competition | AI4SDG3 Challenge With IRCAI Top100 Asia

Innovators representing IRCAI's Top 100 Asia take to the stage with their AI4SDG3 Challenge presentations. [3min]

12:25PM - 12:55PM

10X Stage

Junior Pitch Competition | AI4SDG3 Challenge High School Student Category

High school students take to the stage to present their AI4SDG3 Challenge submissions. [3min]

12:55PM - 1:10PM

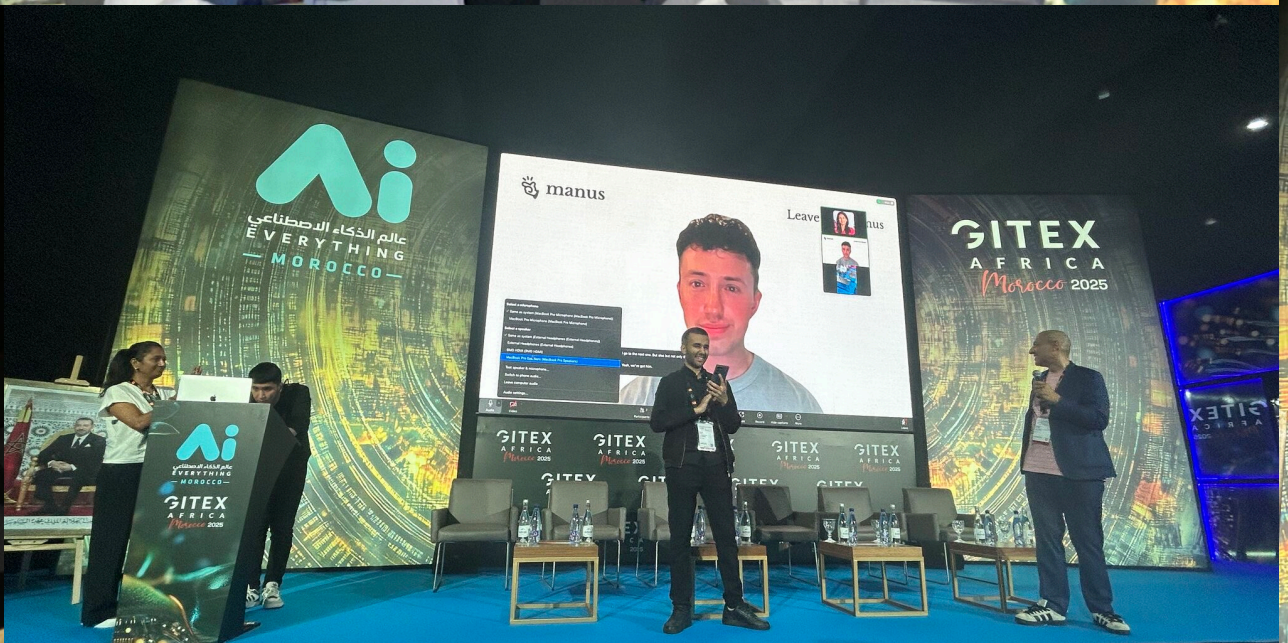
10X Stage

11:30AM -

10X Stage


AFRICA
OUR TIME
NOW

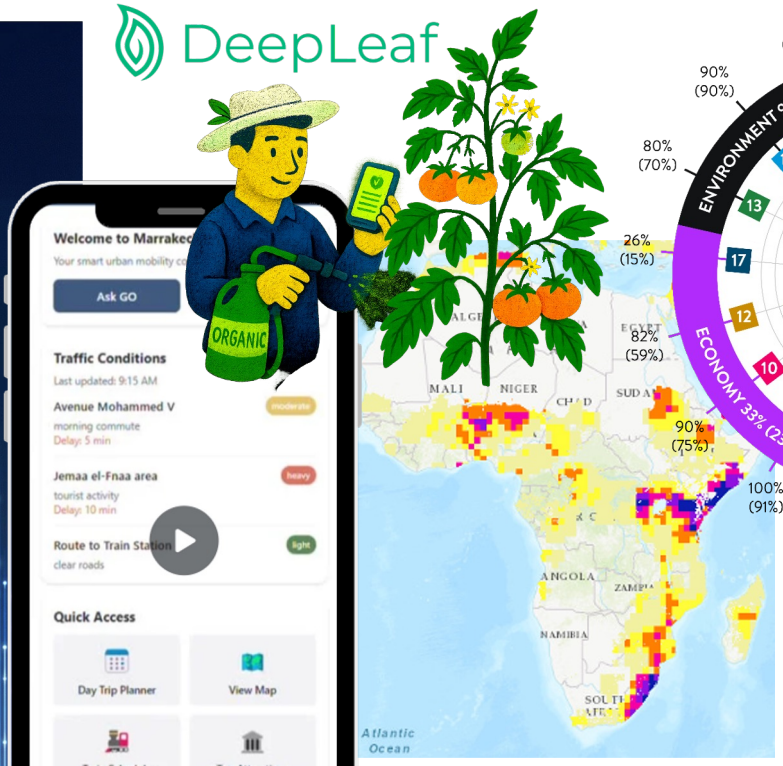

عالم الذكاء الاصطناعي
EVERYTHING
MOROCCO
RENEE CUMMINGS
UNIVERSITY OF VIRGINIA, SCHOOL OF DATA SCIENCE
KENZA GREFFT-ALAMI CEO INJAZ AL MAGHRIB
MANAL ELATTIR
WOMENOMICS INSTITUTE
IRC AI
GITEX AFRICA Morocco 2025

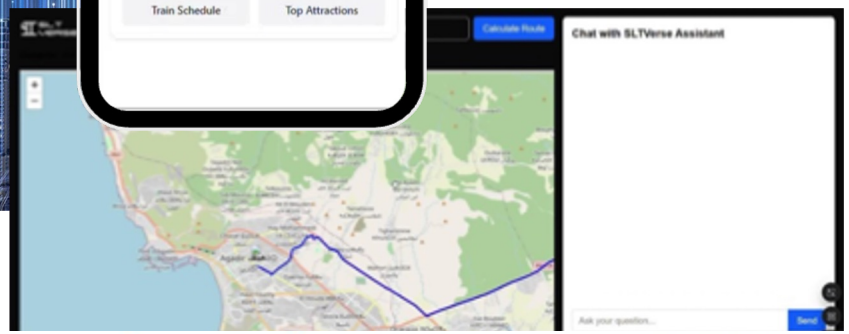

manus
عالم الذكاء الاصطناعي
EVERYTHING
MOROCCO
GITEX AFRICA Morocco 2025

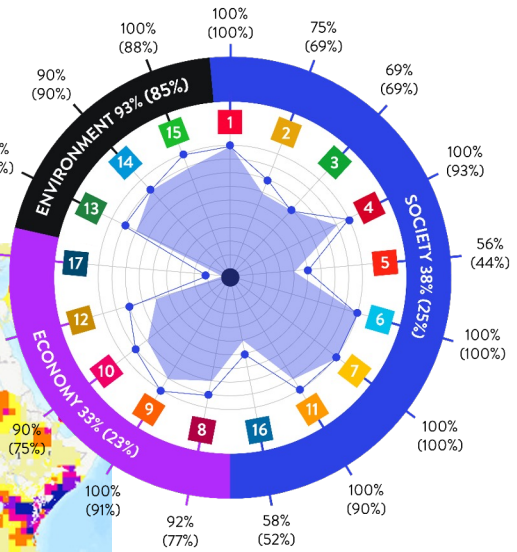

AI4SDG11 CHALLENGE
SUSTAINABLE CITIES & COMMUNITIES
GITEX AFRICA Morocco
ENSIAS
Connected cities, protected communities:
Flood resilience in Africa
A federated AI system built by and for Africa
SOBEK
عالم الذكاء الاصطناعي
EVERYTHING
MOROCCO
GITEX AFRICA Morocco 2025

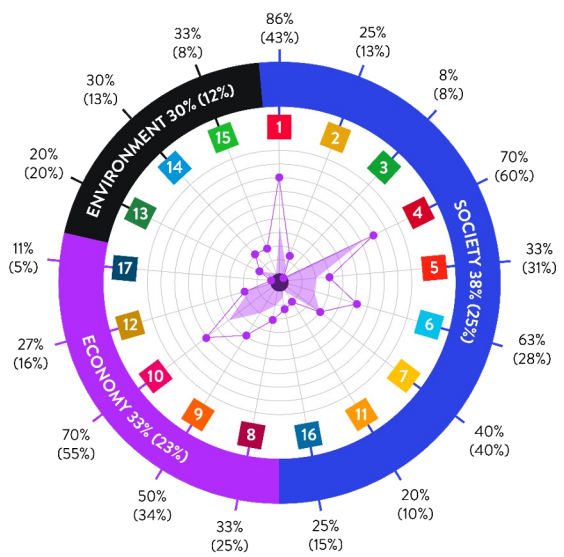
1. **Science Knowledge Extraction for Good Practices**
2. **Data Visualization Narratives on Smart Water**
3. **Appropriate application of LLMs**
4. **SDG6-focused classifiers**
5. **Early detection of floods based on social media**
6. **Looking for trends with AI**
7. **Automated sentiment analysis on media**
8. **Causality analysis on water topics from social media**
9. **Analysis of waterborne disease on scientific articles**
10. **Bias, Ethics & Responsible AI**

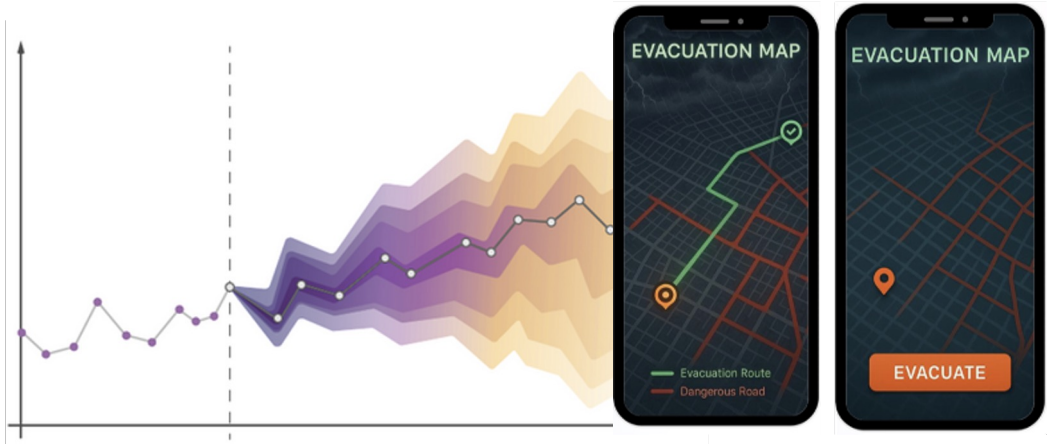
SDG Agentic AI #AI4SDG











https://www.nature.com/articles/s41467-019-14108-y

GITEX AFRICA Morocco

ELIAS

unesco Centre Under the auspices of UNESCO

NAiXUS

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

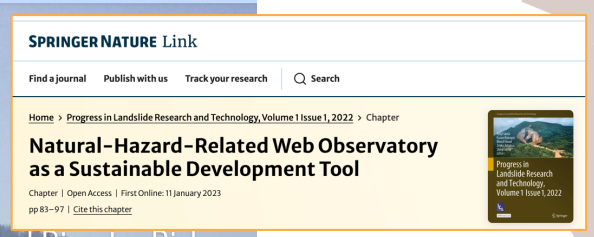
IN AFRICA

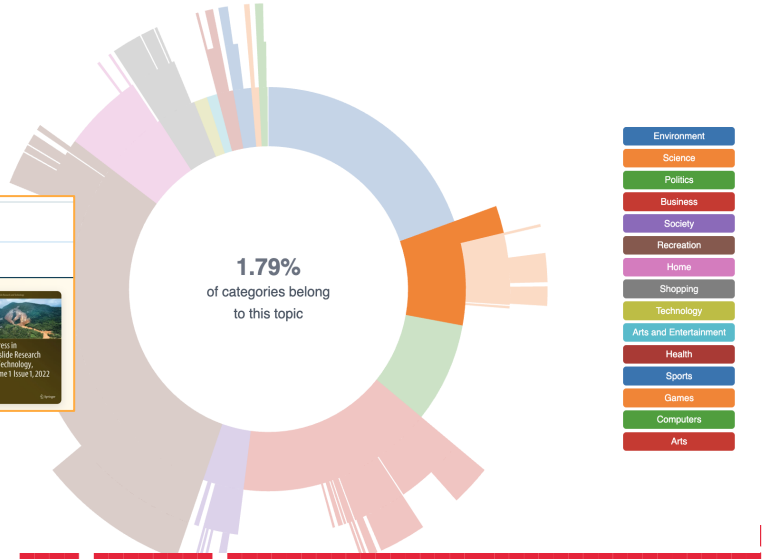
GITEX ASIA Singapore

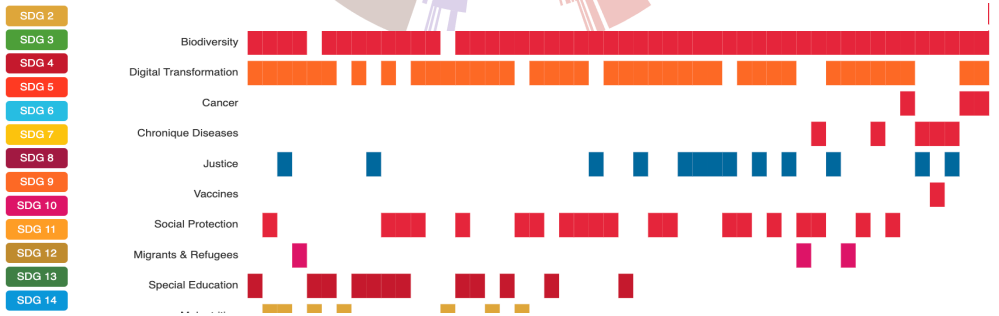
AI EVERYTHING -SINGAPORE-

Disaster Risk Reduction & Landslides

The Landslide Observatory is developed to provide a multifaceted view on landslide-related information worldwide. It includes data from daily multilingual news and published science, as well as yearly ingested statistical indicators, AI policies and open education resources.







https://www.unesco-floods.eu/

ELIAS

unesco Centre Under the auspices of UNESCO

NAiXUS

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

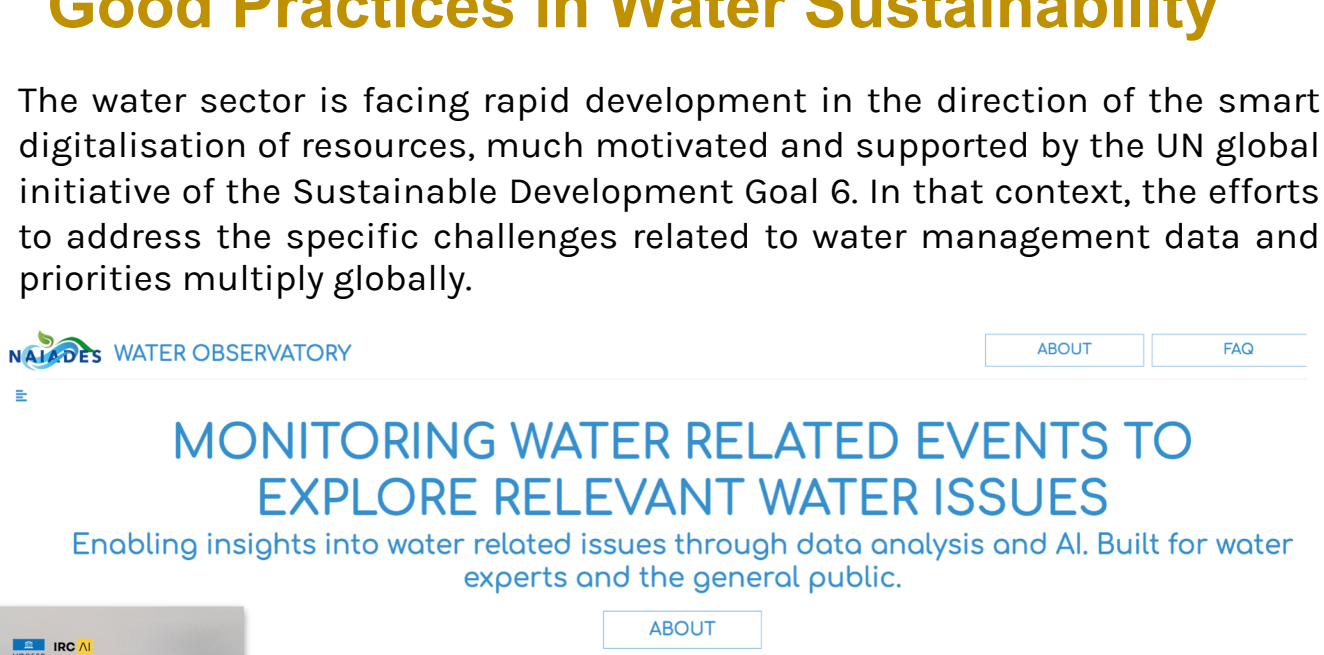
AI IN AFRICA

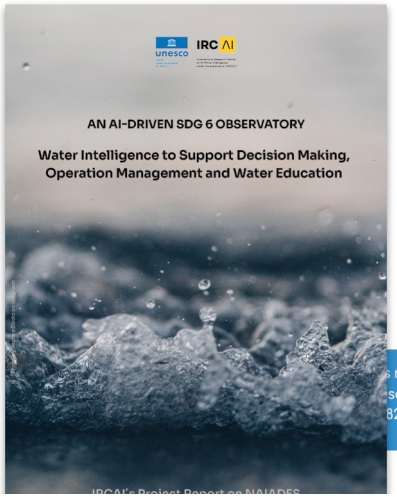
GITEX ASIA Singapore

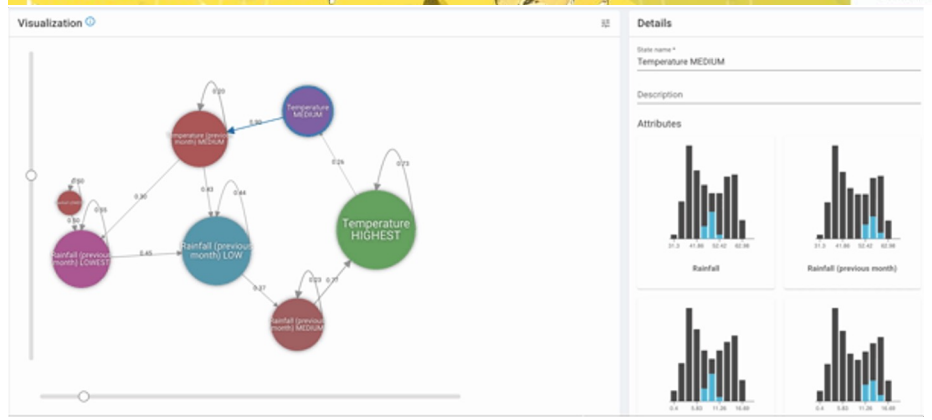
AI EVERYTHING -SINGAPORE-

Good Practices in Water Sustainability

The water sector is facing rapid development in the direction of the smart digitalisation of resources, much motivated and supported by the UN global initiative of the Sustainable Development Goal 6. In that context, the efforts to address the specific challenges related to water management data and priorities multiply globally.







ELIAS

unesco
Centre
Under the auspices of UNESCO

NAiXUS

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

IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

AI Impact on Sustainable development

WORLD NEWS

Canarias24Horas.com - Tenerife será sed
Wed Jul 10 2024, 11:33

OPINION - How can Türkiye make Developme
Wed Jul 10 2024, 11:00

Features Invest in fertil
Wed Jul 10 2024, 05:59

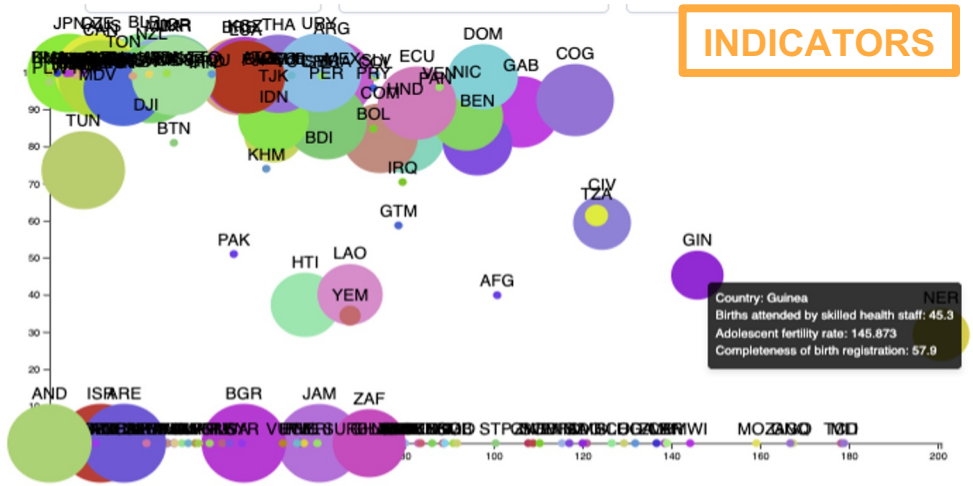
One Health Approach Can Help Ease Climat
Wed Jul 10 2024, 07:28

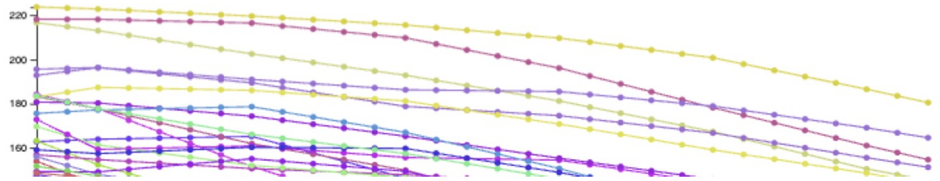
Trilateral Powerhouse: India, Indonesia,
Tue Jul 9 2024, 16:25

LG USA RENEWS PARTNERSHIP WITH TENNESSEE
Tue Jul 9 2024, 15:57

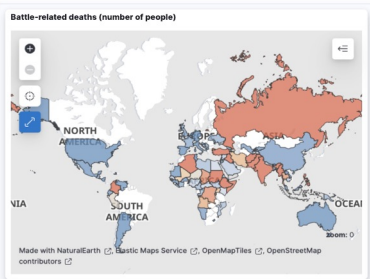
Virtusa and Quality Clouds Forge Strateg

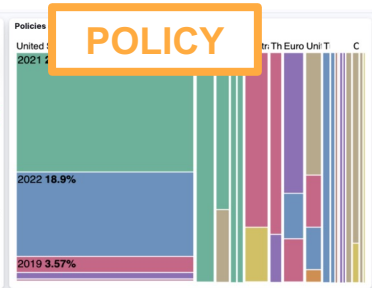
INDICATORS

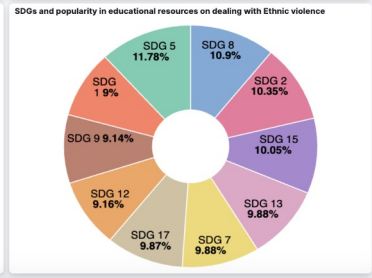




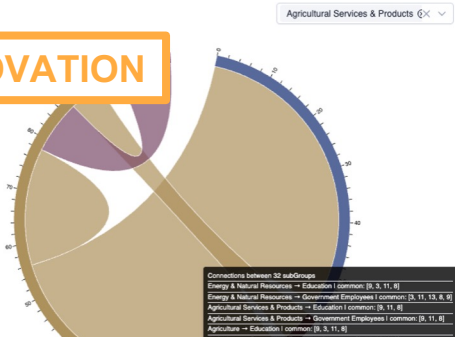
POLICY





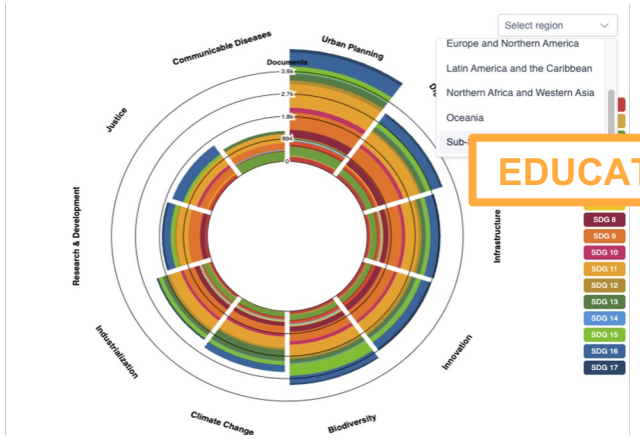


INNOVATION



SCIENCE

EDUCATION



ELIAS

unesco
Centre
Under the auspices
of UNESCO

NAiXUS

IRCAI
International Research Centre
of Artificial Intelligence
under the auspices of UNESCO

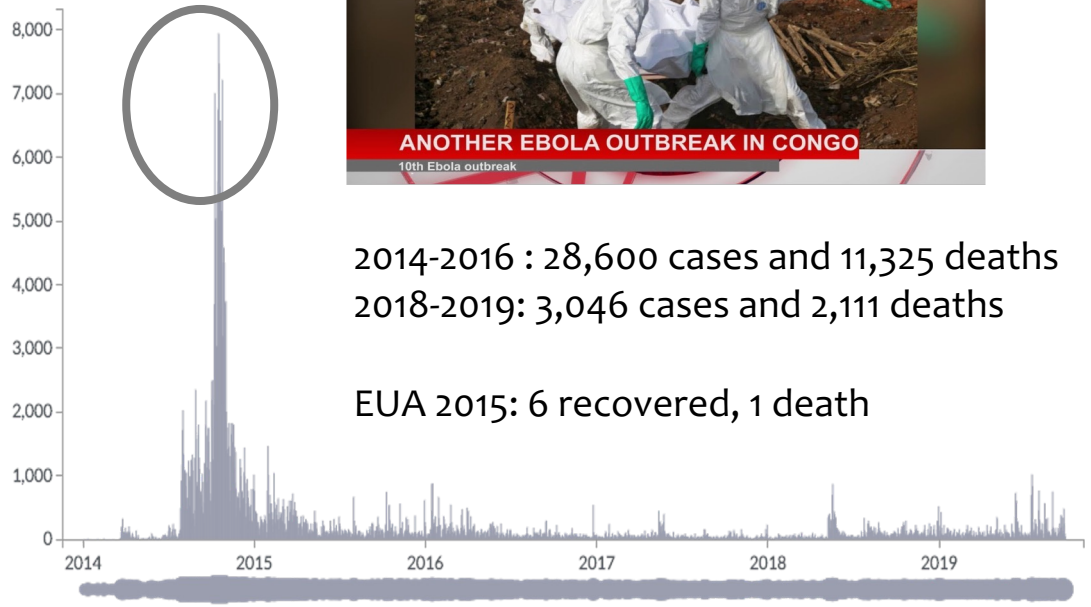
AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

Bias in News on Science

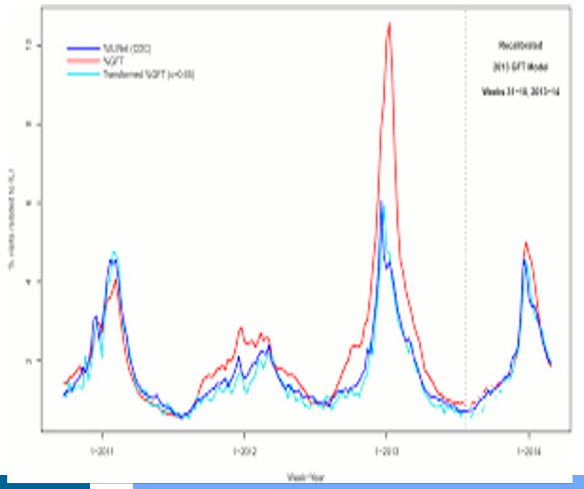
1



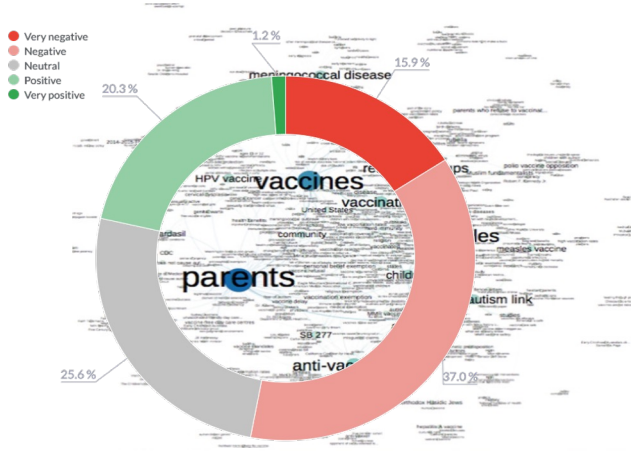
2014-2016 : 28,600 cases and 11,325 deaths
2018-2019: 3,046 cases and 2,111 deaths

EUA 2015: 6 recovered, 1 death

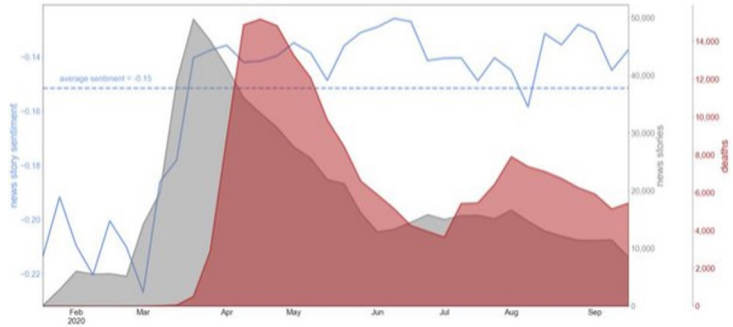
2



3



COVID-19 media coverage sentiment vs. deaths



INFORMATION COMPLEXITY

MANIPULATION / IDEOLOGY

INNACCURACY / MISUNDERSTANDING

ORIGIN LOCATION/ LANGUAGE

ELIAS

unesco
Centre
Under the auspices
of UNESCO

NAiXUS

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

AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

Road to COP30: AI MetAmazonia

Real Florest

20,000 hectares that bring real information to the Platform

Social Impact

Improving the lives of more than 450 families

Reducing the Ecological Footprint

2.2 million tons of carbon to be mitigated

Partnerships with Universities

UNIVERSITY OF CAMBRIDGE

Imperial College London

UNITED NATIONS CLIMATE CHANGE CONFERENCE

COP30 AMAZÔNIA

CUIDAR DO PLANETA PARA O FUTURO DA HUMANIDADE

ELIAS

unesco Centre Under the auspices of UNESCO

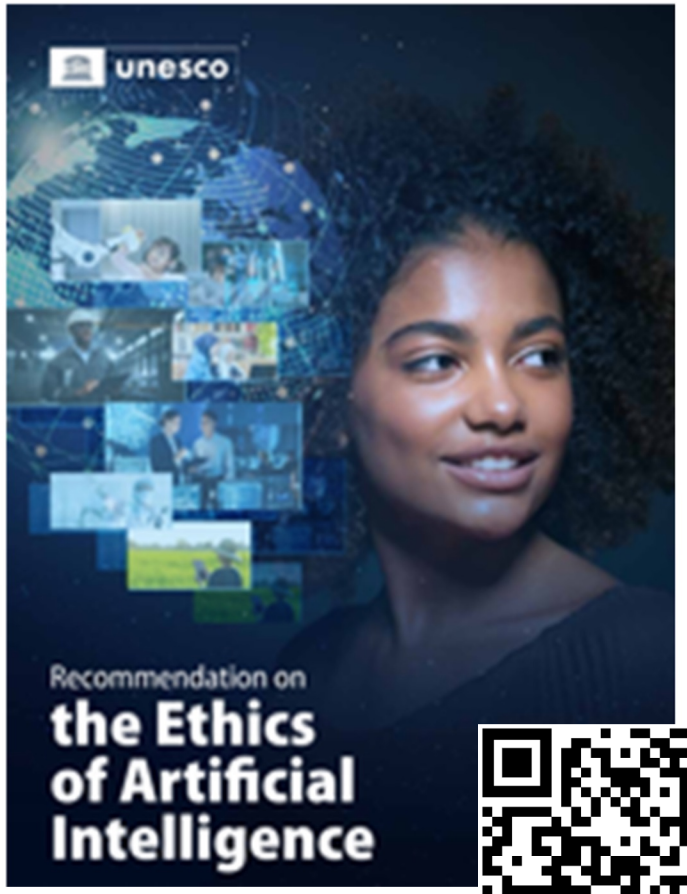
NAiXUS

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

IN AFRICA

GITEX ASIA Singapore

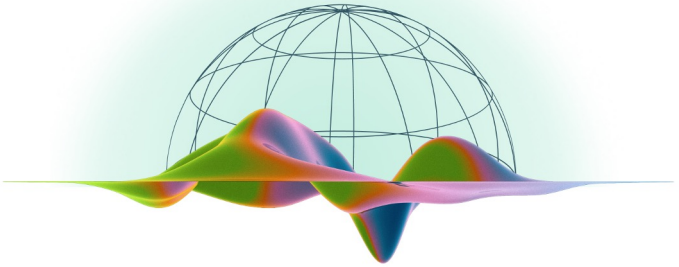
AI EVERYTHING -SINGAPORE-



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)



ELIAS

unesco Centre Under the auspices of UNESCO

NAiXUS

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

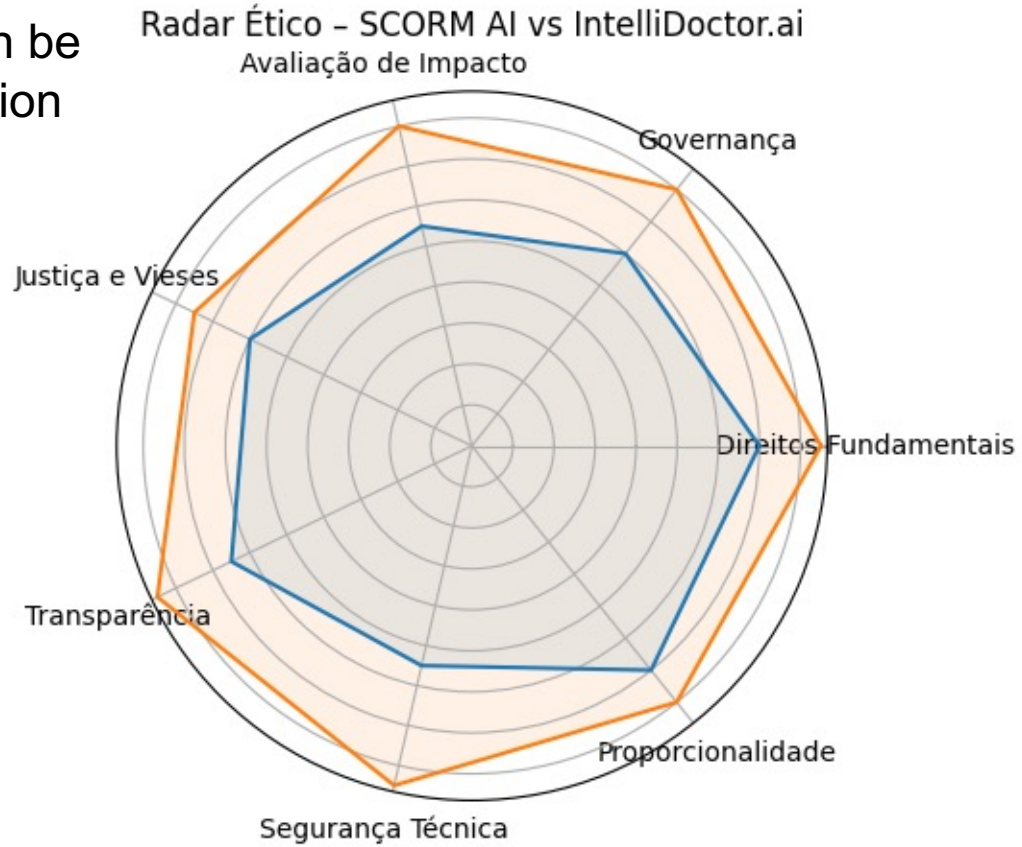
IN AFRICA

GITEX ASIA Singapore

Ai EVERYTHING -SINGAPORE-

AI REGULATORY SANDBOX & AI ETHICS

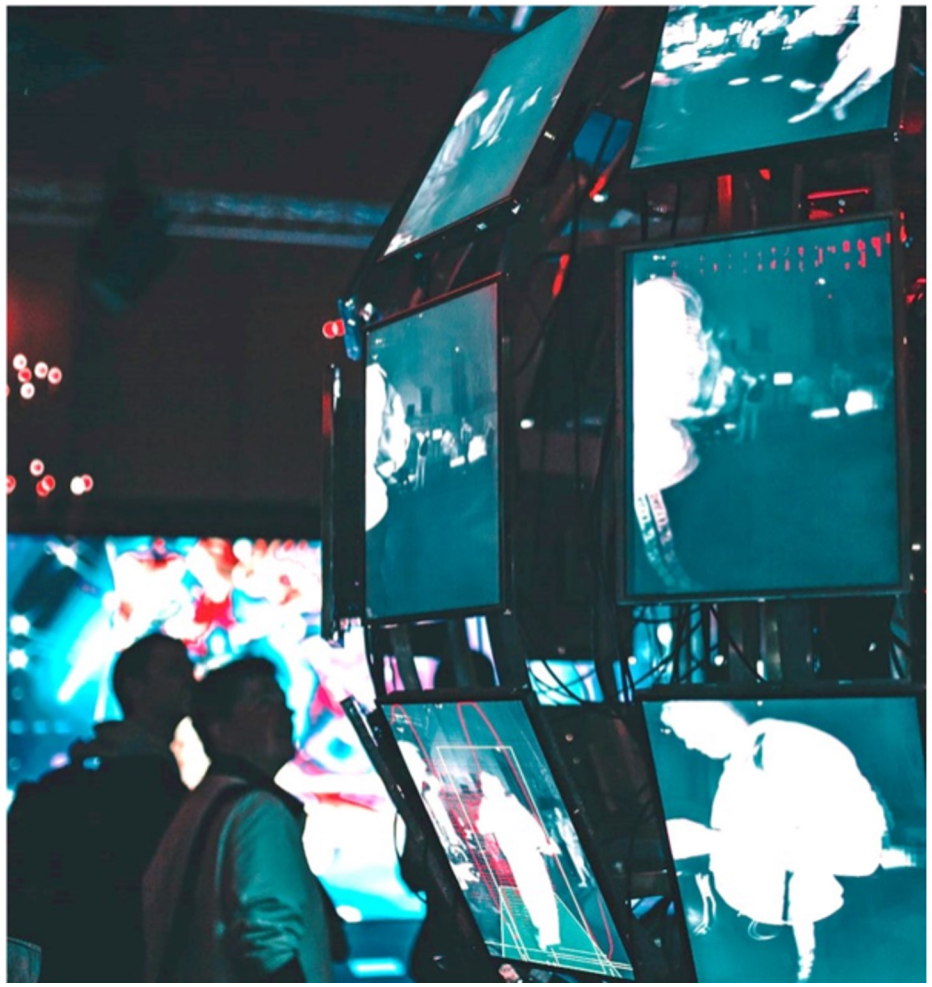
AI regulatory sandbox is a controlled environment established by competent authorities where AI systems can be developed, tested, and validated under regulatory supervision before being placed on the market or put into service.



1. New to market and complying to current regulation
2. Centrality of human dignity and fundamental rights
(Maximum priority – criterion for remaining in the sandbox)
3. Clear governance and accountability
(Without governance → it is not an ethical sandbox)
4. Impact assessment (continuous)
(Central element of the sandbox according to UNESCO)
5. Justice, non-discrimination, and algorithmic bias
6. Transparency, explainability, and communication
7. Technical safety and robustness
8. Proportionality and necessity

Framework for AI Explainability, Transparency & Trustworthiness

ELIAS unesco Centre Under the auspices of UNESCO NAiXUS IRC AI International Research Centre of Artificial Intelligence under the auspices of UNESCO IN AFRICA GITEX ASIA Singapore AI EVERYTHING -SINGAPORE-



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

ELIAS

unesco Centre Under the auspices of UNESCO

NAiXUS

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

AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

Human Rights and AI Systems Interaction

Human Rights Embedded in Functioning of the AI Systems

AI Systems Can Violate Human Rights

AI Systems Can Contribute to Respect of Human Rights

ELIAS

unesco
Centre
Under the auspices
of UNESCO

NAiXUS

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

AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

Human Rights Embedded in Functioning of the AI Systems

Right to Privacy and Data Protection

- Unauthorized data collection
- AI surveillance without adequate oversight
- Data breaches exposing sensitive data

Right to Non-Discrimination

- Algorithmic biases that can lead to unfair treatment of vulnerable groups (excluding or targeting)
- Predictive policing or access to public services

Right to a Fair Trial and Right to an Effective Remedy

- Lack of transparency in algorithmic decisions
- Decision of these systems are unclear and cannot be tested

ELIAS

unesco Centre Under the auspices of UNESCO

NAiXUS

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

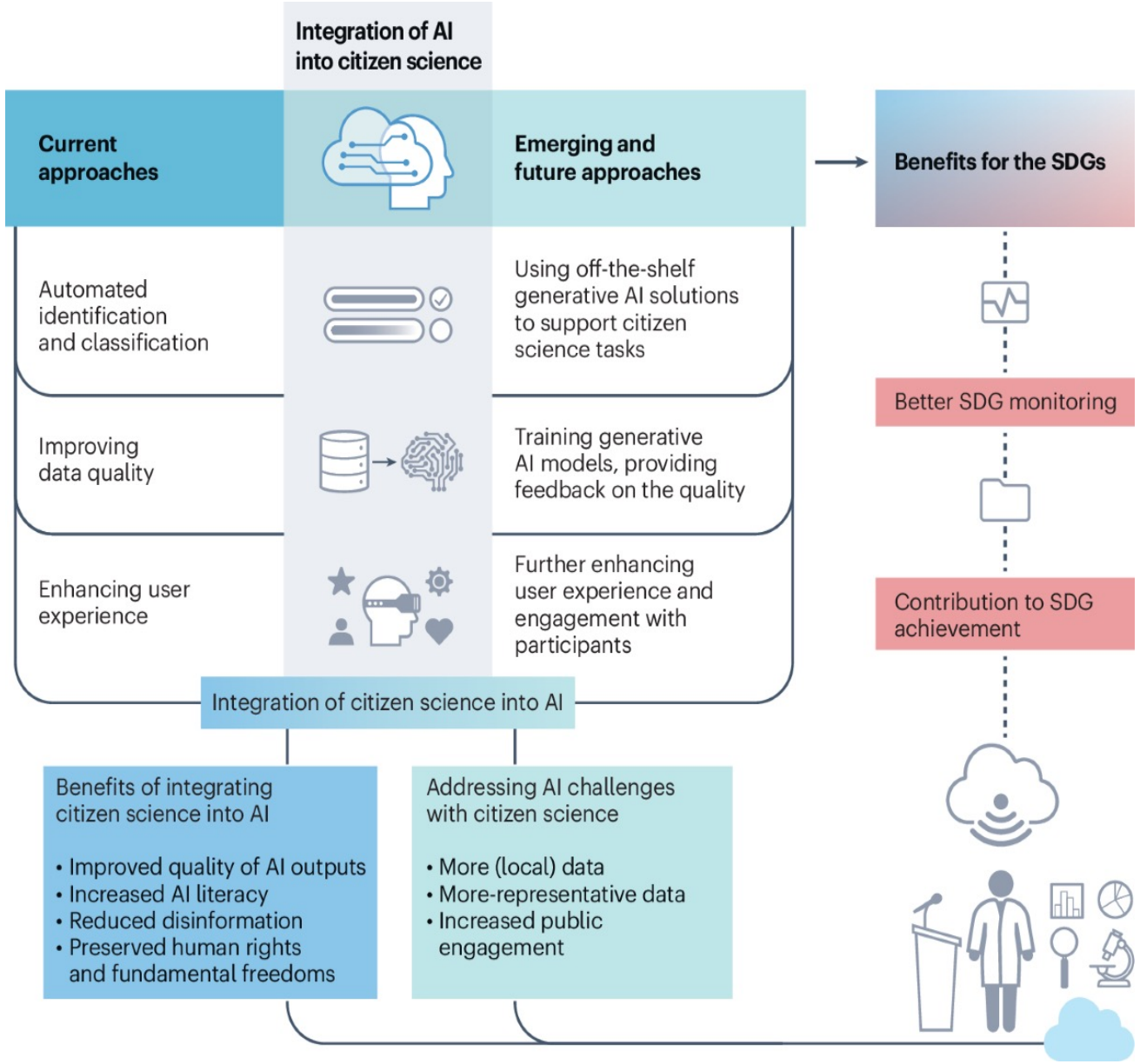
AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

https://unesdoc.unesco.org/ark:/48223/pf0000385082

Leveraging the collaborative power of AI and citizen science for sustainable development, Nature Sustainability 2024



ELIAS

unesco Centre Under the auspices of UNESCO

NAiXUS

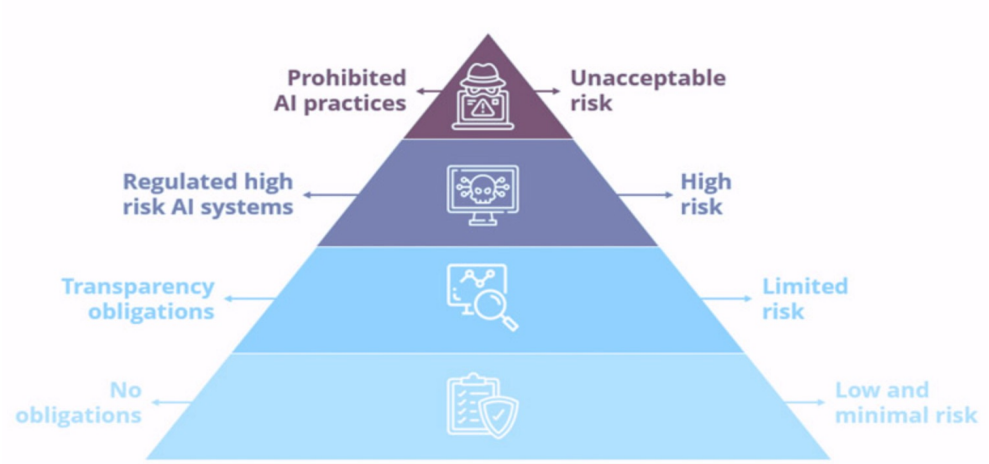
IRC AI International Research Centre of Artificial Intelligence under the auspices of UNESCO

AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

- 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.
  -



EU AI ACT

unesco
Centre
Under the auspices of UNESCO

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

ELIAS

ELIAS

unesco
Centre
Under the auspices of UNESCO

NAiXUS

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

AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

- 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 NET

HumanE AI Net:

The HumanE AI Network

ELIAS

unesco Centre Under the auspices of UNESCO

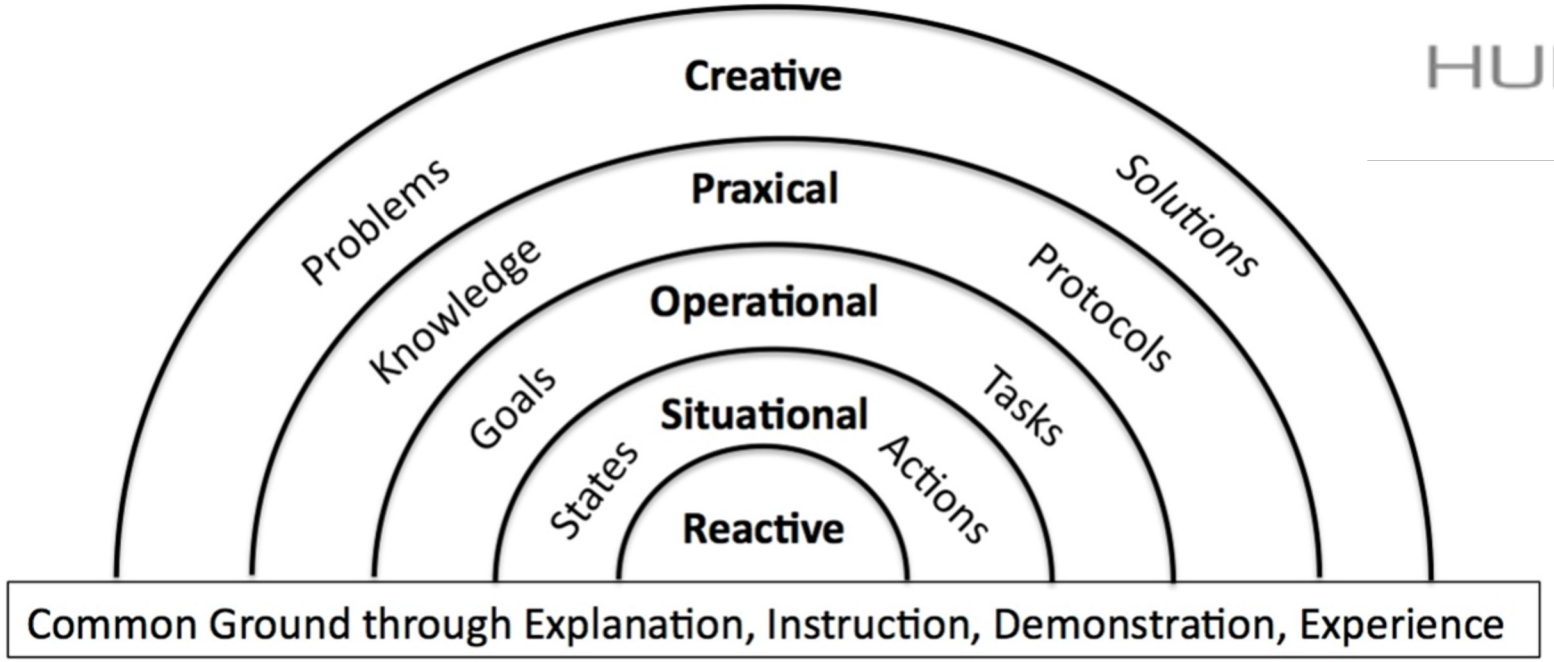
NAiXUS

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

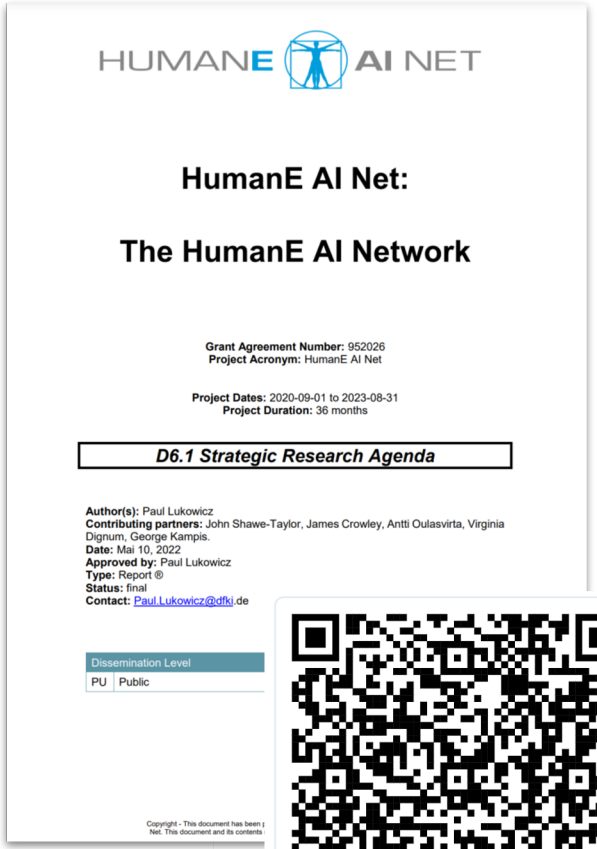
AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-



HUMANE AI NET


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

Copyright - This document has been p
Net. This document and its contents

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

ELIAS

unesco Centre Under the auspices of UNESCO

NAiXUS

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

AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

- Fairness: several measures that are not equivalent
- Bias: AI has revealed bias in the data
- Privacy: techniques for ensuring personal data is not revealed through models developed by AI systems

An intelligent person

An unintelligent person

A very competent person

A very incompetent person

Images created with DALL-E via Bing Image Creator

ELIAS

unesco Centre Under the auspices of UNESCO

NAiXUS

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

IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

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

ELIAS

unesco
Centre
Under the auspices
of UNESCO

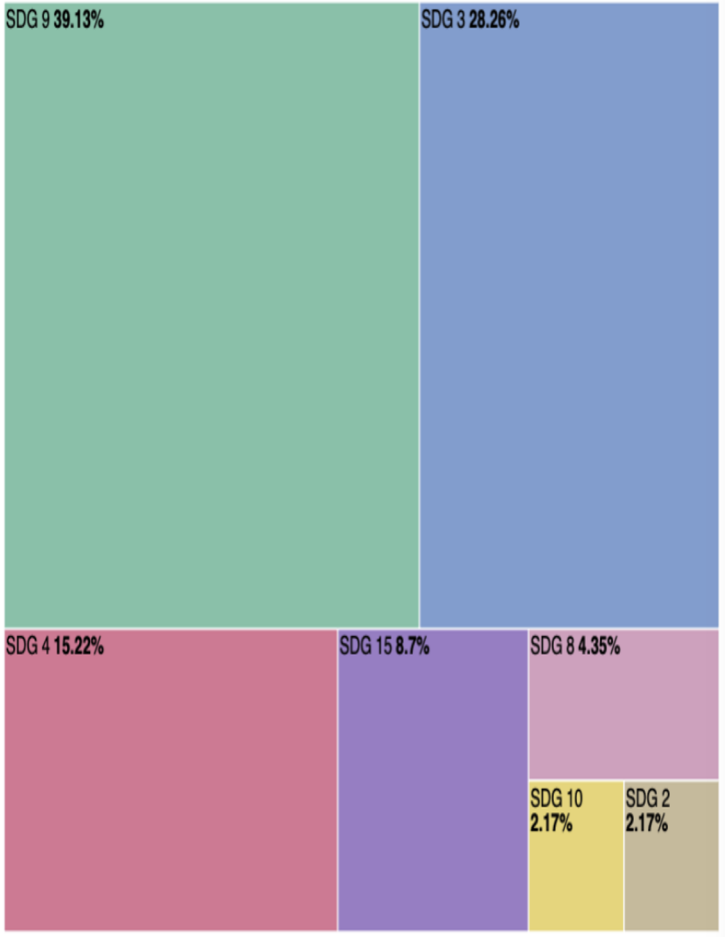
NAiXUS

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

AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-



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.

image: https://www.youtube.com/watch?v=riyh_CIshTs

Backend Structure Overview

```
backend/
├── api/                    # API endpoints
│   ├── routes.py           # Main API routes
│   └── admin_routes.py     # Admin API routes
├── models/                 # Data models
│   ├── research.py         # Research-related mc
│   └── admin.py            # Admin-related model
├── services/               # Service layer
│   └── llm_service.py      # LLM interaction ser
├── utils/                  # Utility functions
│   ├── file_manager.py     # File operations
│   ├── admin_utils.py      # Admin utilities
│   ├── auth.py             # Authentication util
│   └── logger.py           # Logging utilities
├── config/                 # Configuration
│   └── config.py           # Application configu
├── storage/                # Storage for files a
│   ├── prompts/            # Prompt templates
│   ├── outputs/            # Generated research
│   └── logs/               # System logs
├── main.py                 # Application entry p
├── requirements.txt        # Python dependencies
├── README.md               # Documentation
├── Dockerfile              # For containerizatic
└── .env.example            # Environment variabl
```

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.

ELIAS

unesco Centre Under the auspices of UNESCO

NAiXUS

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

AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

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.

ELIAS

unesco
Centre
Under the auspices
of UNESCO

NAiXUS

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

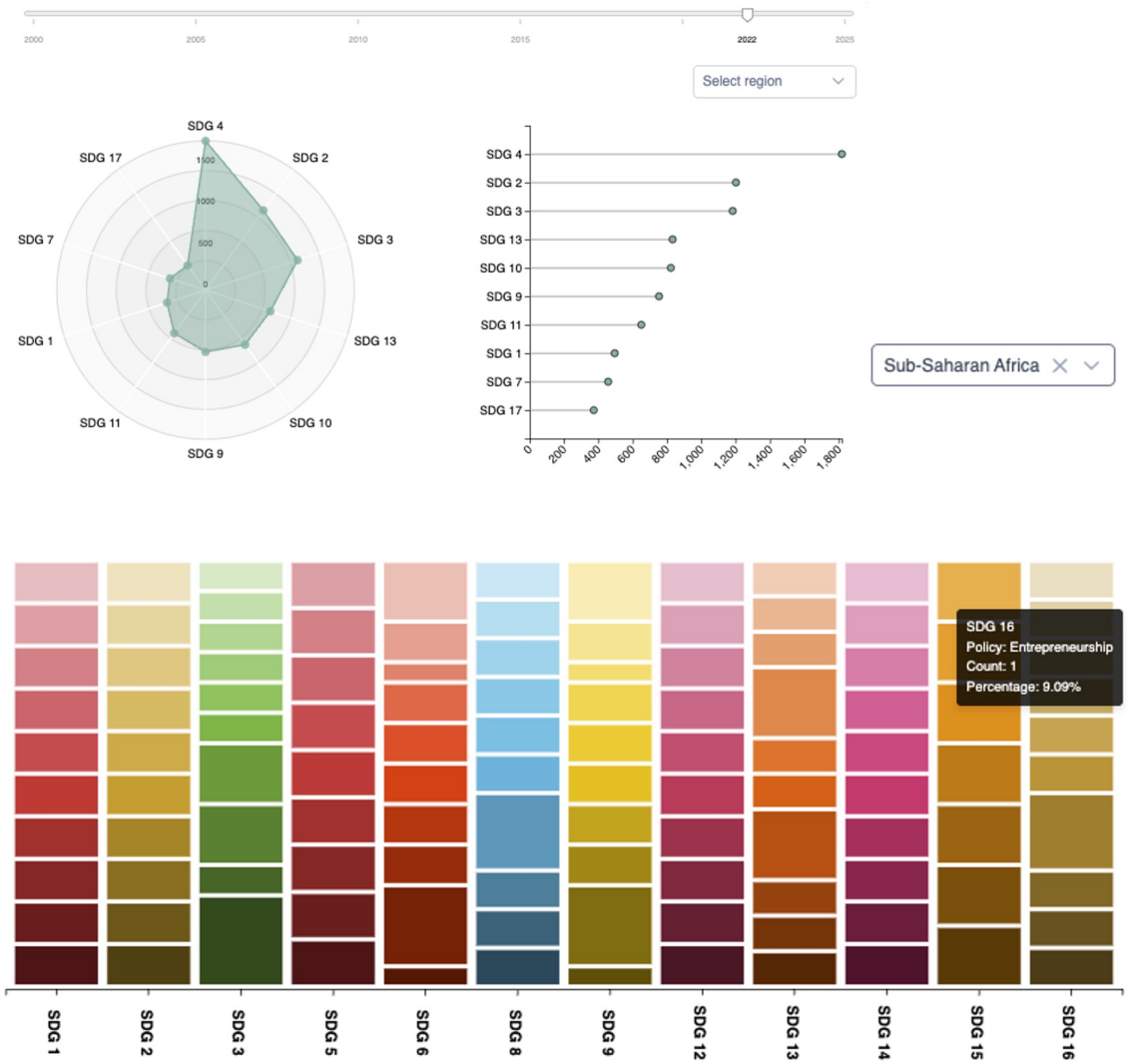
AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

- ➔ 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

ELIAS

unesco Centre Under the auspices of UNESCO

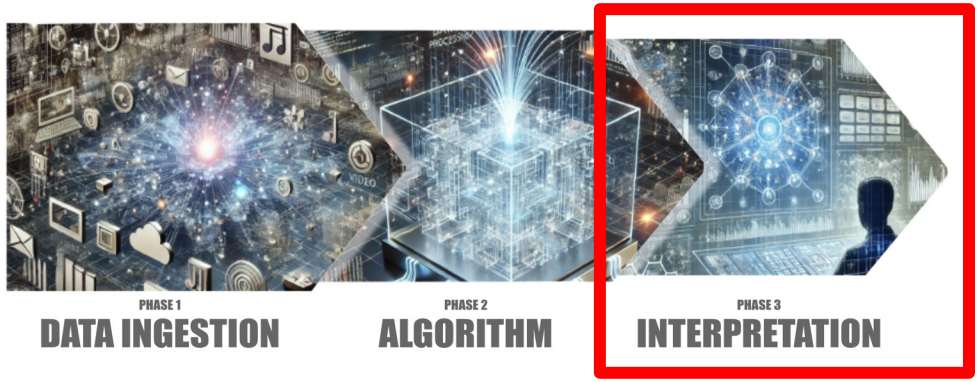
NAiXUS

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

AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-



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.

ELIAS

unesco
Centre
Under the auspices
of UNESCO

NAiXUS

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

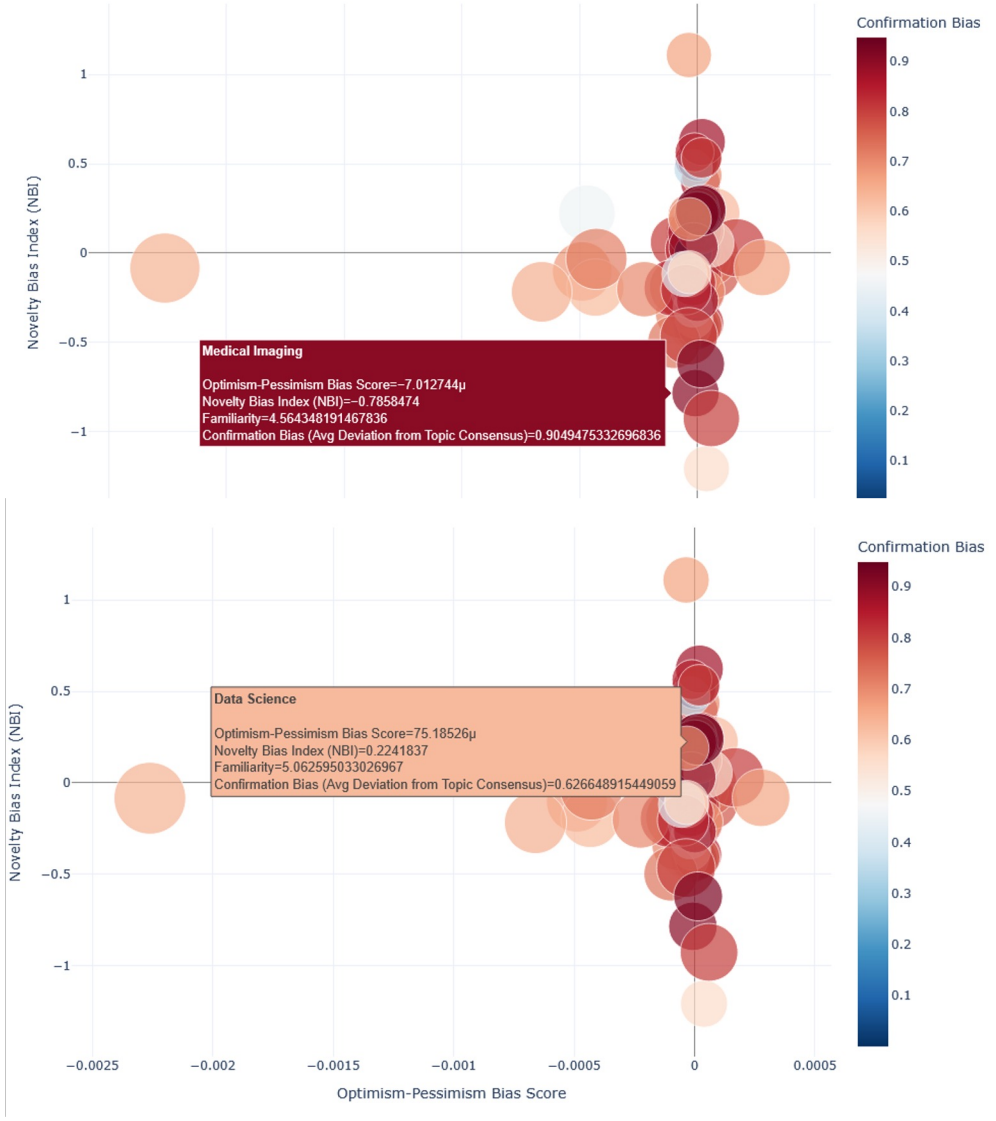
IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

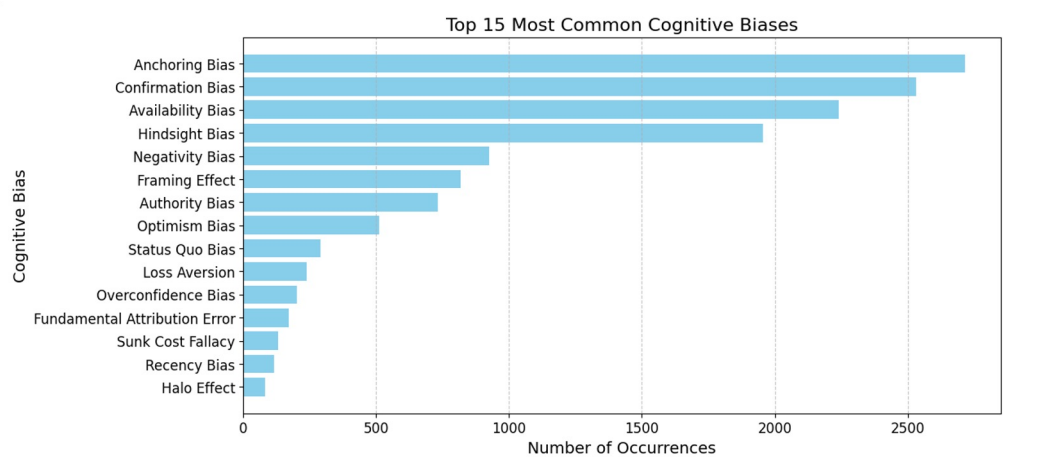
- Integration of cognitive science with computational social science.
- Empirical framework for bias detection in AI discourse.
- Use of semantic linking to align media content with AI taxonomy.
- Identification of cognitive mechanisms shaping AI adoption narratives.
- First large-scale framework connecting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

Cognitive Bias Aware.

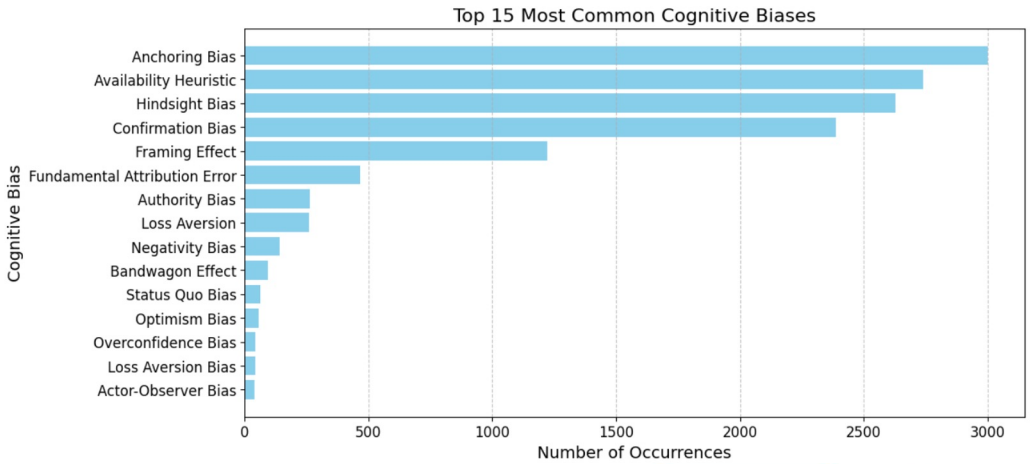


elias-ai.eu 101120237

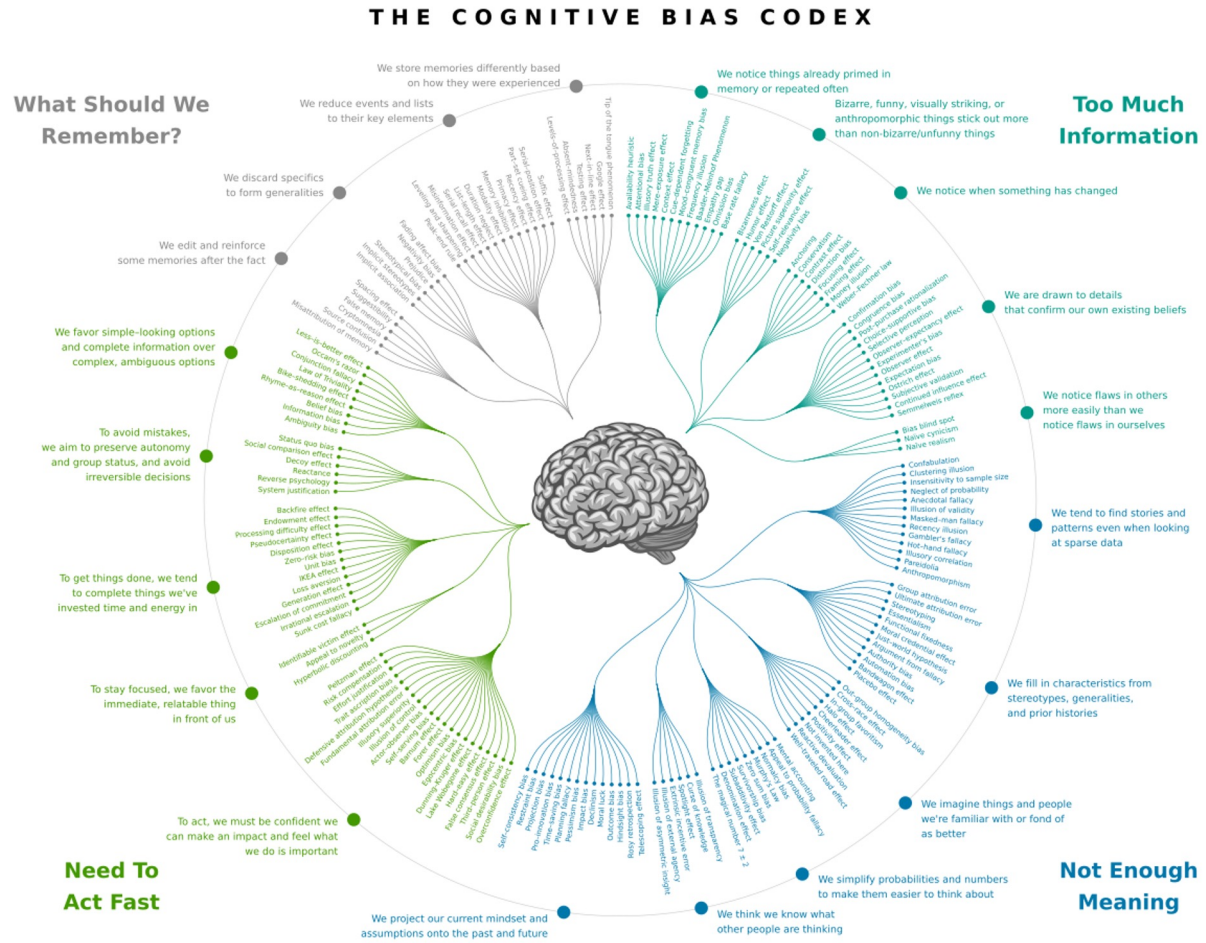
ELIAS: Cognitive Bias on News



DEEPSEEK
Topic: self-driving cars

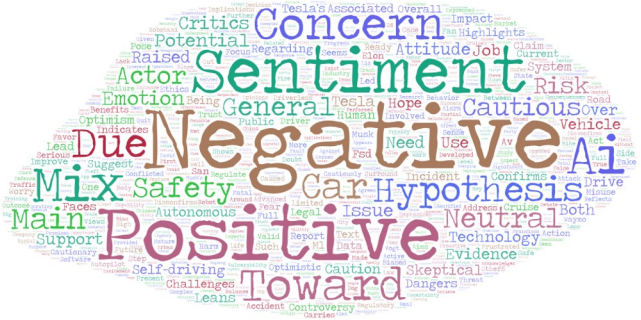


LLAMA
Topic: self-driving cars



ELIAS
unesco Centre Under the auspices of UNESCO
NAiXUS
IRCAI International Research Centre of Artificial Intelligence under the auspices of UNESCO
AI IN AFRICA
GITEX ASIA Singapore
AI EVERYTHING -SINGAPORE-

ELIAS: AI Cognitive Bias



Data: incidents from news related to technology use (linked to the OECD AI ontology and AI Event Detection)

- exploring data by source
- exploring data by time
- exploring data by location

from emotions and sentiment to bias (**confirmation**: related to emotions, sentiment, source and time, **proximity**: related to source and time, **negativity**: related to emotions, sentiment, source and time aspect)

Also, we will improve some prompting techniques for understanding of cognitive biases in news

ELIAS: AI Cognitive Bias

Example: self driving cars

source	negative	positive	neutral	other	count	% negativ	% postitiv	% neutral	% other
forbes.com	8	9	15	17	49	0.16	0.18	0.31	0.35
dailymail.co.uk	11	3	11	7	31	0.35	0.10	0.35	0.23
finance.yahoo.com	8	5	2	4	19	0.42	0.26	0.11	0.21
theregister.com	9	5	2	2	18	0.50	0.28	0.11	0.11
techtimes.com	7	4	2	3	16	0.44	0.25	0.13	0.19
teslarati.com	7	4	1	3	15	0.47	0.27	0.07	0.20
jalopnik.com	5	0	2	7	14	0.36	0.00	0.14	0.50
electrek.co	5	3	3	3	14	0.36	0.21	0.21	0.21
marketscreener.co	4	4	2	2	12	0.33	0.33	0.17	0.17
news.yahoo.com	6	3	2	1	12	0.50	0.25	0.17	0.08
futurism.com	5	0	4	2	11	0.45	0.00	0.36	0.18
theguardian.com	3	2	0	6	11	0.27	0.18	0.00	0.55
sfist.com	4	3	1	3	11	0.36	0.27	0.09	0.27
benzinga.com	2	2	4	2	10	0.20	0.20	0.40	0.20
nbcbayarea.com	8	0	0	2	10	0.80	0.00	0.00	0.20
businessinsider.co	3	4	2	1	10	0.30	0.40	0.20	0.10
washingtonpost.co	5	2	2	1	10	0.50	0.20	0.20	0.10
techcrunch.com	6	2	0	2	10	0.60	0.20	0.00	0.20
wonderfulenginee	3	1	3	2	9	0.33	0.11	0.33	0.22
ca.sports.yahoo.co	0	3	4	2	9	0.00	0.33	0.44	0.22
carscoops.com	4	5	0	0	9	0.44	0.56	0.00	0.00
autoevolution.com	5	1	0	3	9	0.56	0.11	0.00	0.33
abc7news.com	2	2	2	3	9	0.22	0.22	0.22	0.33

on average, more **negative** sentiment towards hypothesis related to self-driving cars

possible confirmation bias

source	date	sentiment llm	sentiment classic
futurism.com	5/30/2023	neutral	-0.098039216
futurism.com	6/9/2023	The sentiment towards the hypothesis is negative.	0.066666667
futurism.com	7/10/2023	negative	0.098039216
futurism.com	9/16/2023	The hypothesis carries a negative sentiment highlighting potential safety risks associated with Musk's decisions on sensor technology for autonomous driving.	-0.011764706
futurism.com	10/4/2023	cautious	-0.341176471
futurism.com	10/22/2023	Mixed	-0.215686275
futurism.com	10/27/2023	negative	0
futurism.com	11/7/2023	Cruise's safety performance around children has proved to be inadequate contradicting the main hypothesis.	0
futurism.com	11/9/2023	neutral	-0.105882353
futurism.com	11/16/2023	The sentiment towards the main hypothesis is negative.	-0.129411765
futurism.com	1/29/2024	The sentiment towards the main hypothesis is mixed with some perceiving it as a safety improvement and others as inadequate.	-0.090196078

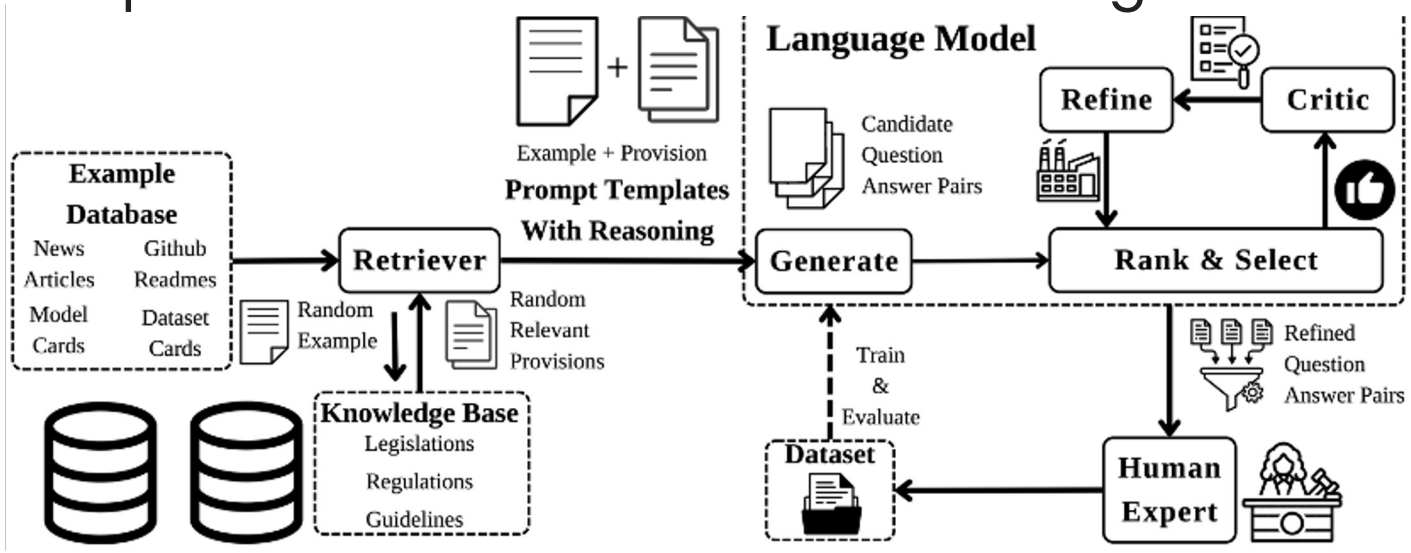
no confirmation bias

source	date	sentiment llm	sentiment classic
theguardian.com	4/20/2022	positive	0.066666667
theguardian.com	8/7/2022	Skeptical	0.043137255
theguardian.com	5/16/2023	The sentiment towards the hypothesis is negative as it involves the theft of sensitive technology.	0.152941176
theguardian.com	7/7/2023	The general sentiment towards the hypothesis seems to be in favor of the safety and environmental concerns raised by driverless cars.	0.356862745
theguardian.com	7/26/2023	Concerned about risks and dominance of self-driving cars supportive of public transportation	0.019607843
theguardian.com	8/10/2023	positive	-0.160784314
theguardian.com	8/14/2023	negative	-0.184313725
theguardian.com	10/31/2023	The general sentiment towards the hypothesis is controversial due to conflicting evidence.	0
theguardian.com	11/8/2023	cautiously optimistic	-0.121568627
		The sentiment towards the hypothesis is cautious optimism as the focus seems to be on	

ELIAS: AI Algorithmic Fairness

Create an LLM assistant that can check an AI system for compliance with provided legal & ethical guidelines (incl. fairness)

- Starting with the dataset
  - Generating examples for annotation
- An algorithm for compliance text indexing and querying that preserves structure while retrieving relevant provisions



Preparing AI for Compliance: Initial Steps of a Framework for Teaching LLMs to Reason About Compliance

Makovec, Rei, and Novalija

RuleML+RR'24: Companion Proceedings of the 8th International Joint Conference on Rules and Reasoning (To appear)

ELIAS/DATAPACT: AI Algorithmic Fairness

DATAPACT

Indexing

1. Split document into pre-defined **articles** and **paragraphs**
2. Further segment paragraphs into overlapping **passages**
3. Encode (**Embed**) each passage using a dense retrieval embedding model

Querying

1. Encode the query using the same model
2. Compute (dot product) similarity between query and passage **embeddings**
3. Retrieve top-*k* **passages** with scores above threshold *t*
4. Map retrieved passages back to their respective **articles**

Prompt Generation

1. Use a predefined template incorporating up to *n* **articles**
2. Prompt the ***LLM + "Let's think step by step"***

Consider the following articles of legislation, provided between triple backticks,
and nothing else:
```EU AI Act: Articles: 6, 73, 82, 111, 80, 20, 60, 99``` [text omitted]
Under these articles and only these articles and ignoring those that are not applicable, as a legal compliance expert, answer: what are the implications of the provided articles to the following incident, in triple backticks:
```a 50-year-old man driving on Autopilot was killed when his Tesla plowed under a semi truck.```
Answer article by article and don't provide any additional introduction or conclusions.
Let's think step by step.

ELIAS

unesco Centre Under the auspices of UNESCO

NAiXUS

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

AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

ELIAS: AI Algorithmic Fairness

Example

Article 73: Reporting of Serious Incidents
Should the Autopilot system be classified as a high-risk AI system, the provider (Tesla) would be required to report the incident to the market surveillance authorities within the stipulated time frame if a causal link between the AI system and the accident is established or reasonably suspected.

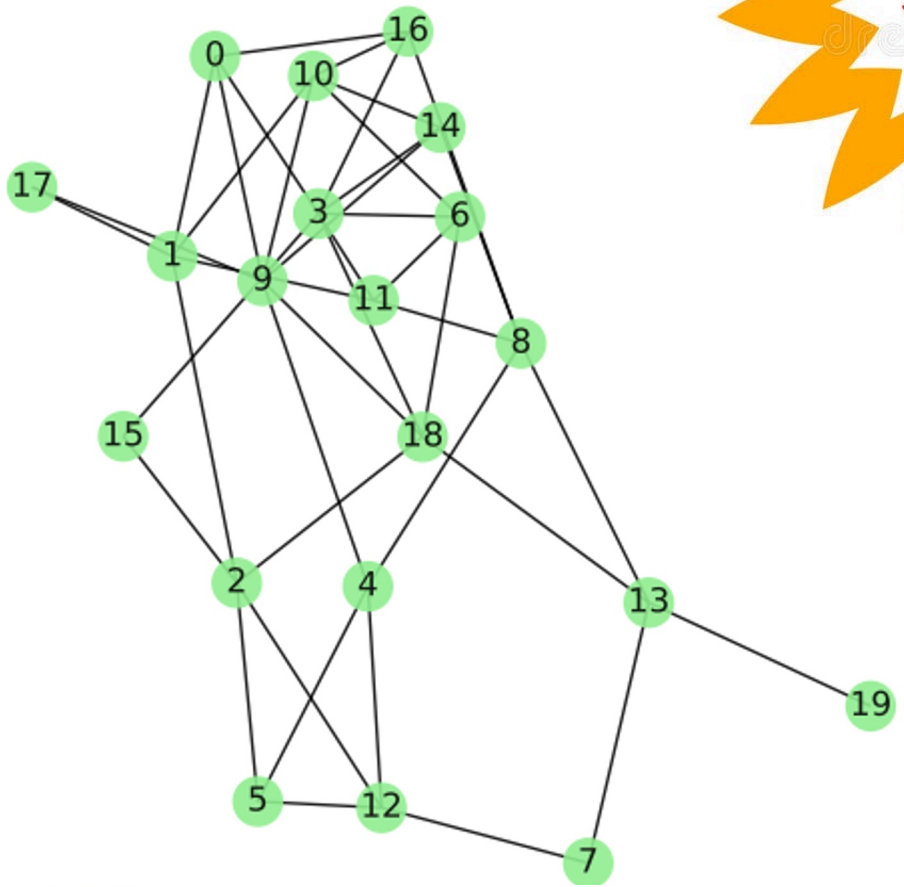
Article 82: Compliant AI Systems Which Present a Risk
If after an evaluation, it is found that the Autopilot system, despite compliance with regulations, presents a risk to safety, health, or fundamental rights, the market surveillance authority could mandate Tesla to take appropriate measures to mitigate the risk.

Article 80: Procedure for Dealing with AI Systems Classified by the Provider as a Not High-Risk in Application of Annex III
If Tesla had classified the Autopilot system as not high-risk and an evaluation by the market surveillance authority finds otherwise, Tesla would be required to comply with high-risk AI system regulations, including taking necessary corrective actions.

Article 60: Testing of High-Risk AI Systems in Real World Conditions Outside AI Regulatory Sandboxes
If the Autopilot system is under testing as a high-risk AI system, specific conditions and approvals are required for

ELIAS: Multi Agent Learning

- Regret minimization with N agents over a communication network G
- Agents can only talk to their neighbors (no message passing)
- A random number of agents is active at each time step
- We study the extent to which communication compensates for the partial feedback
- Special cases: full information (experts) and bandit feedback



N. Cesa-Bianchi, T. Cesari, R. Della Vecchia.
Cooperative Online Learning with Feedback Graphs.
Submitted, 2024.

RAIDO

raido-project.eu

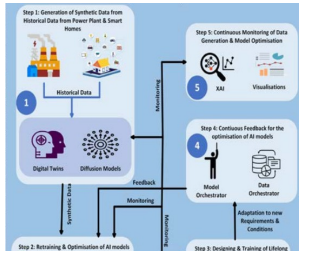
Reliable AI and Data Optimization

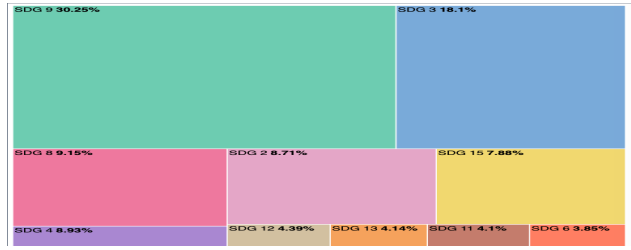
An integrated platform for Green and Energy-efficient data and model related operatio

Green AI Orchestration Bridging Trustworthy AI and Edge AI through tinyML for Frugal Intelligence

Joao Pita Costa[a,*], Marco Zennaro[b], Ioana Ntinou[c,1], John Shawe-Taylor[a]

[a] *International Research Centre on AI under the auspices of UNESCO, Jamova cesta 39, Ljubljana, 1000, Slovenia*





Continuous Compliance with Legislations for Privacy, IRP & Data Collection

AI Ethics & Princples by Design Framework

Reinforced Benchmarking & Feedback-based Progress Monitoring

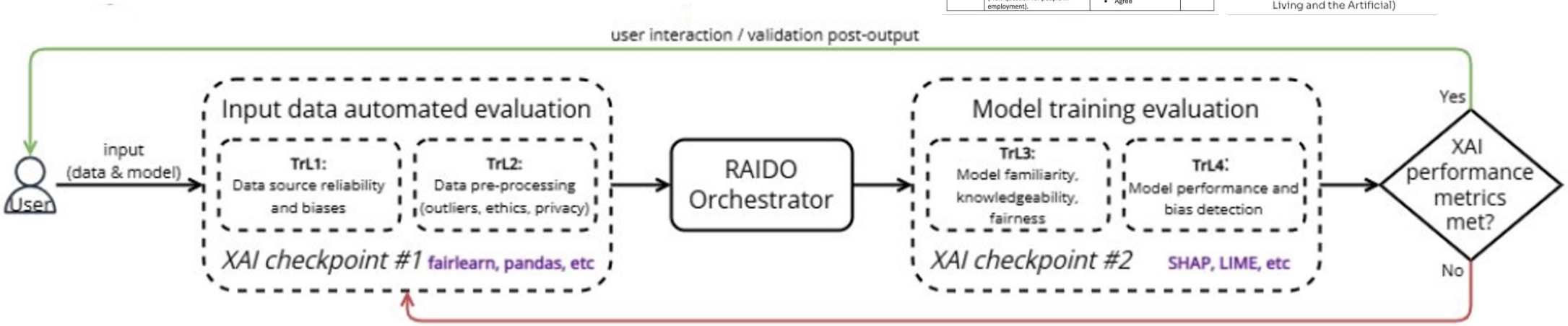
Acceptance of AI use in life and job/industry			
20	I am interested in exploring new technological developments for my field of work/in my daily life.	• Strongly Disagree • Disagree • Neutral/Neither Agree nor Disagree • Agree • Strongly Agree	
21	I could imagine including AI in my current work or in my daily life.	• Strongly Disagree • Disagree • Neutral/Neither Agree nor Disagree • Agree • Strongly Agree	
22	I see the benefit of using AI tools in my current work or in my daily life.	• Strongly Disagree • Disagree • Neutral/Neither Agree nor Disagree • Agree • Strongly Agree	
23	Policymakers strongly support AI.	• Strongly Disagree • Disagree • Neutral/Neither Agree nor Disagree • Agree • 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 • Disagree • Neutral/Neither Agree nor Disagree • Agree • 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 • Disagree • Neutral/Neither Agree nor Disagree • Agree • 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 • Disagree • Neutral/Neither Agree nor Disagree • 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

ELIAS

unesco
Centre
Under the auspices of UNESCO

NAiXUS

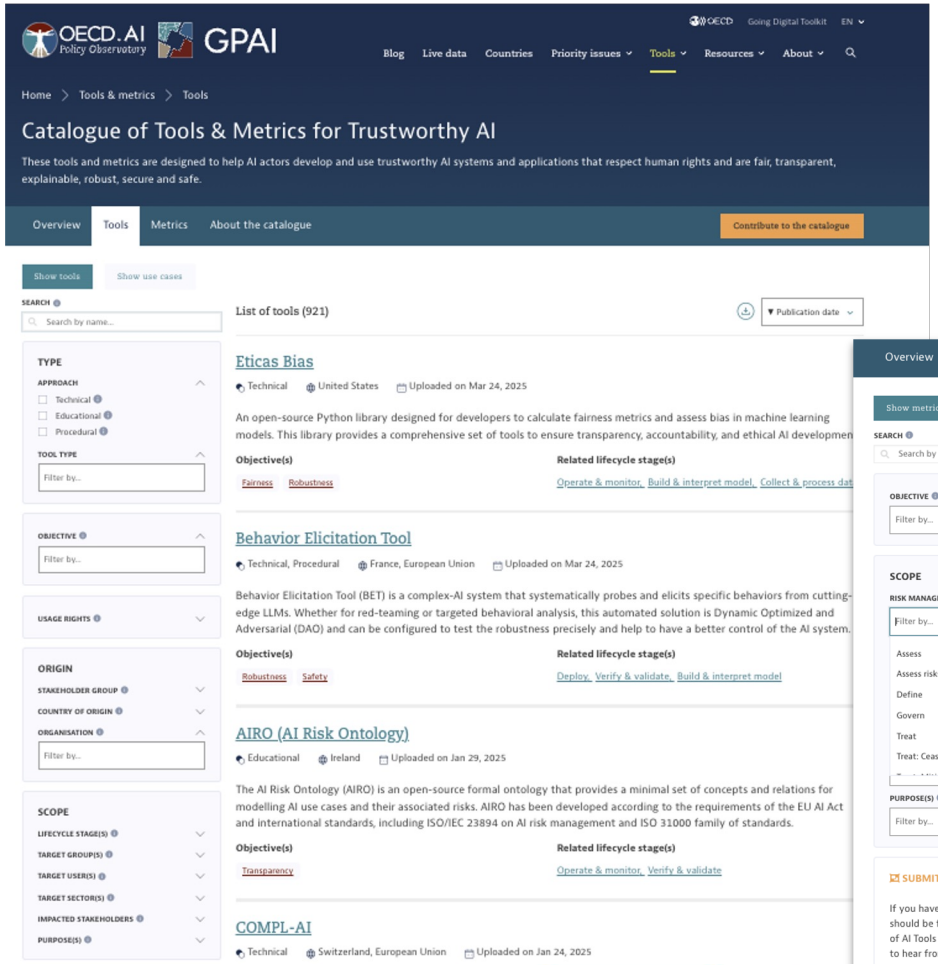
IRC AI
International Research Centre of Artificial Intelligence under the auspices of UNESCO

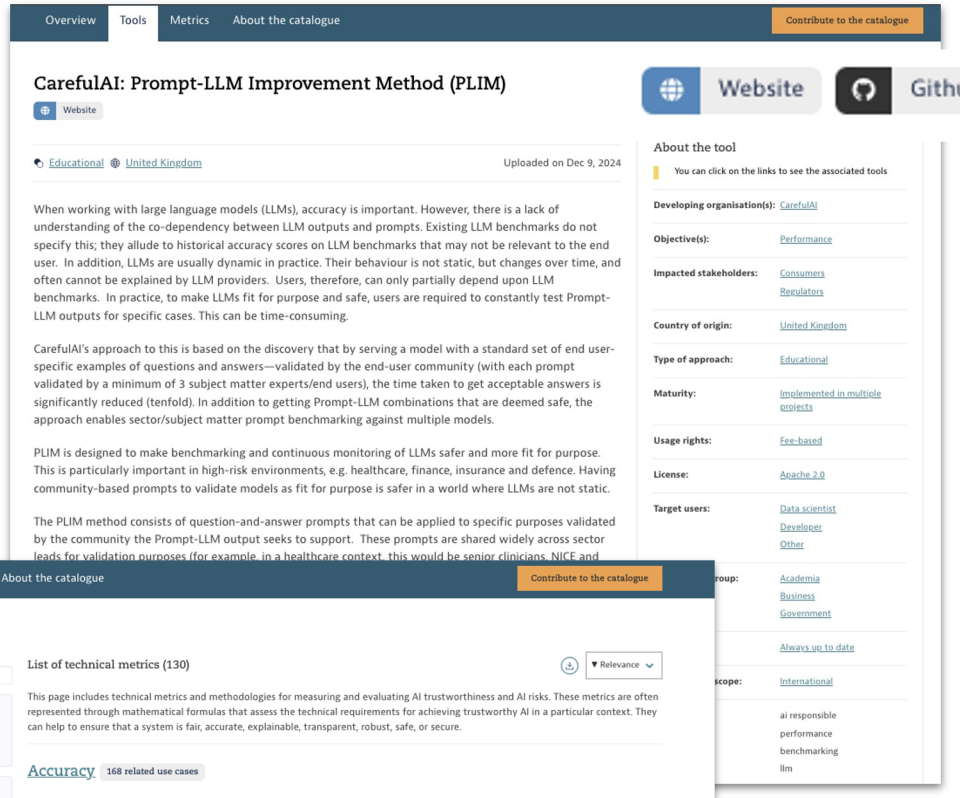
IN AFRICA

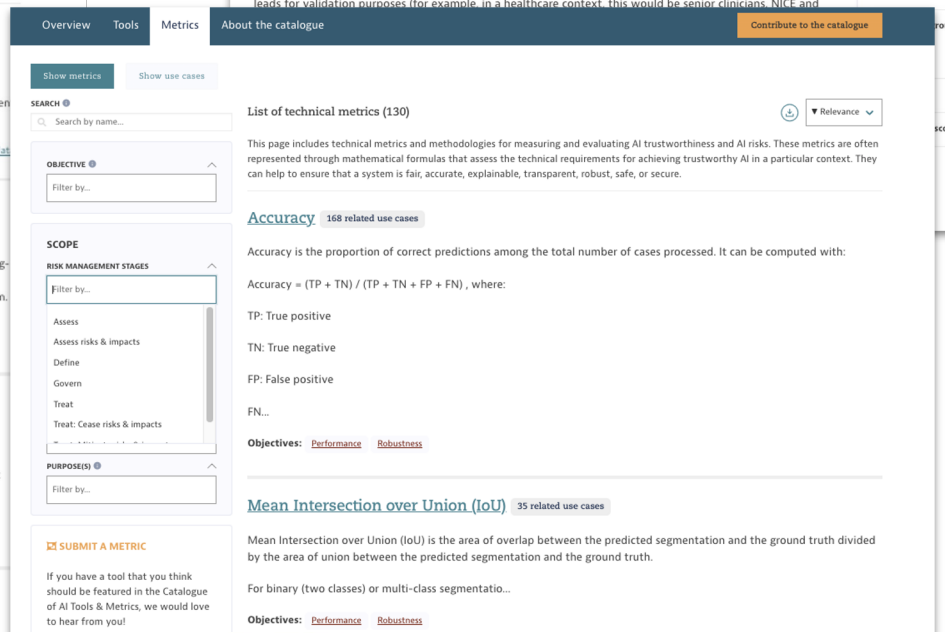
GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

OECD Bias Tools







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

ELIAS

unesco Centre Under the auspices of UNESCO

NAiXUS

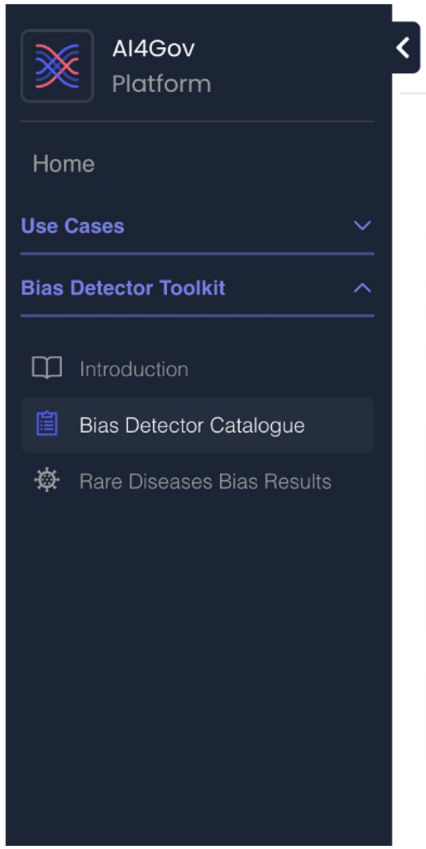
IRC AI International Research Centre of Artificial Inteligence under the auspices of UNESCO

AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

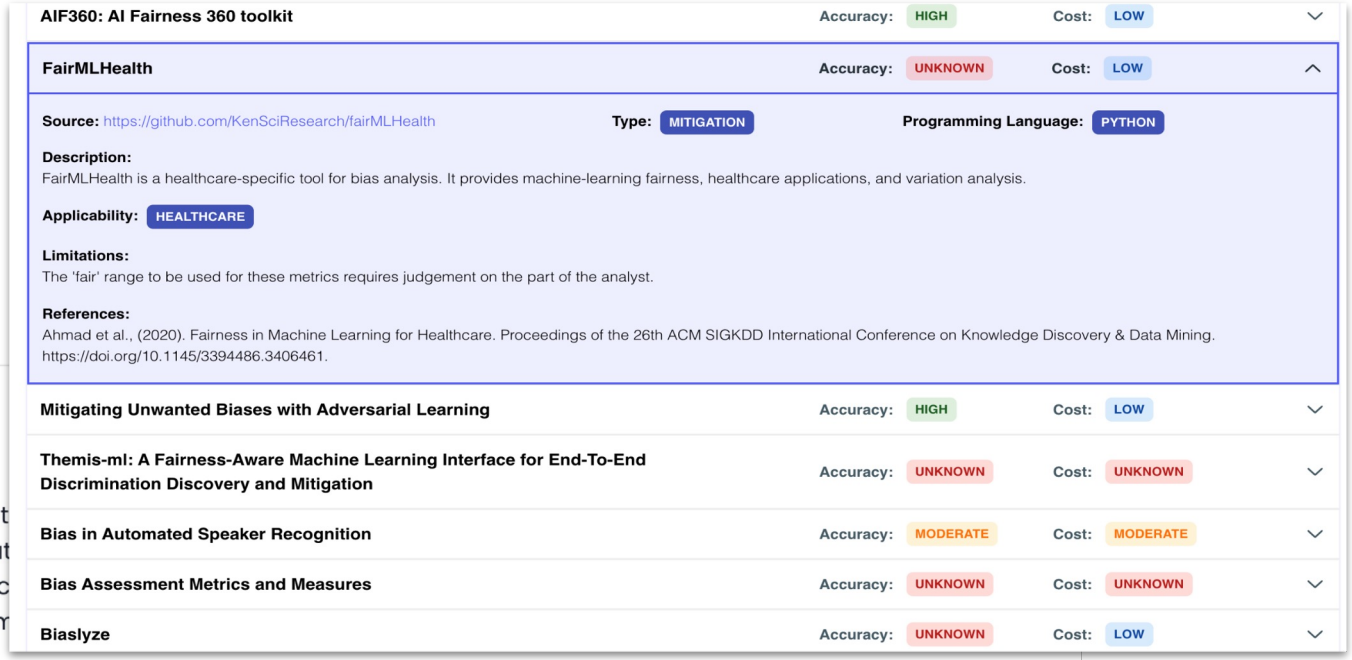
AI4GOV Bias Tools



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



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

AI4GOV MOOC


AI4Gov
Trusted AI for Transparent Public Governance fostering Democratic Values

CERTIFICATE OF COMPLETION

This certificate is awarded to

for succesfully completed the online learning course titled

Trustworthy and Democratic AI - Fundamentals

This course is designed to provide a comprehensive understanding of the foundations of building trustworthy and democratic Artificial Intelligence (AI). From understanding the basics of AI to the ethical considerations that underpin responsible AI, it covers both theoretical concepts and practical applications, with a strong emphasis on addressing bias and ensuring ethical governance.

Course Modules:
1. Introduction to Trustworthy and Democratic AI (Fundamentals)
2. Bias in AI - Understanding Bias
3. Data and Bias
4. Ethical AI Governance

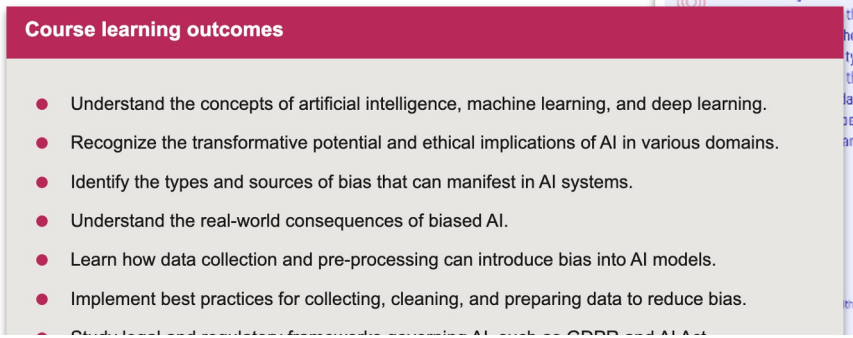
Learning Outcomes:
the concepts of artificial intelligence, machine learning, and deep learning.
he transformative potential and ethical implications of AI in various domains.
types and sources of bias that can manifest in AI systems.
the real-world consequences of biased AI.
lata collection and preprocessing can introduce bias into AI models.
est practices for collecting, cleaning, and preparing data to reduce bias.
nd regulatory frameworks governing AI, such as GDPR and AI Act.

Developed by Jožef Stefan Institute

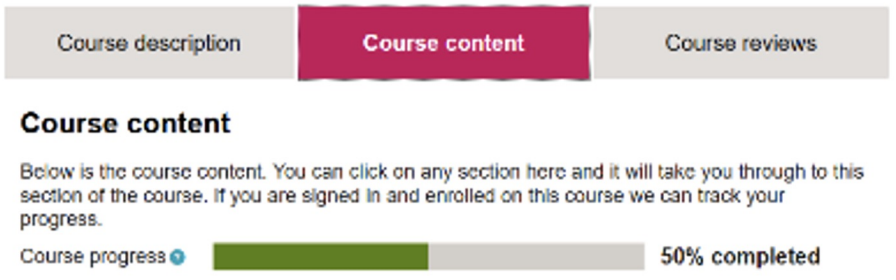
Course link

This course is made available under CC BY-NC-SA 4.0

thin the Horizon Europe project AI4Gov: Trusted AI for Transparent Public Governance fostering Democratic Values (Grant agreement ID: 101094905).


Course learning outcomes

- Understand the concepts of artificial intelligence, machine learning, and deep learning.
- Recognize the transformative potential and ethical implications of AI in various domains.
- Identify the types and sources of bias that can manifest in AI systems.
- Understand the real-world consequences of biased AI.
- Learn how data collection and pre-processing can introduce bias into AI models.
- Implement best practices for collecting, cleaning, and preparing data to reduce bias.
- Study legal and regulatory frameworks governing AI, such as GDPR and AI Act.


Course description | **Course content** | Course reviews

Course content

Below is the course content. You can click on any section here and it will take you through to this section of the course. If you are signed in and enrolled on this course we can track your progress.

Course progress 50% completed

- **Module 1: Introduction to Trustworthy and Democratic AI (Fundamentals)**
- **Module 2: Bias in AI - Understanding Bias**
- **Module 3: Data and Bias**
- **Module 4: Ethical AI Governance**
- **Resources & Authorship**

https://ai4gov-project.eu/home/resources/

ELIAS

unesco Centre Under the auspices of UNESCO

NAiXUS

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

AI IN AFRICA

GITEX ASIA Singapore

AI EVERYTHING -SINGAPORE-

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/