GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

11-12 NOVEMBER 2021 PARIS

13h30-14h30

RESPONSIBLE AI WORKING GROUP

Yoshua Bengio
Mila - Quebec Artificial Intelligence Institute, Canada

Dino Pedreschi
University of Pisa, Italy

Raja Chatila
Sorbonne University, France

Nicolas Miailhe
The Future Society, France

THE RESPONSIBLE AI WORKING GROUP'S MANDATE

The mandate of the Working Group is "foster and contribute to the responsible development, use and governance of human-centred AI systems, in congruence with the UN Sustainable Development Goals".

CLIMATE CHANGE AND AI

Recommendations for Government Action

CENTRE FOR AI & CLIMATE

Climate Change AI

Raja Chatila
Co-Lead
Sorbonne University

Nicolas Mialhe
Co-Lead
Founder & President
The Future Society

RESPONSIBLE AI FOR SOCIAL MEDIA GOVERNANCE

A proposed collaborative method for studying the effects of social media recommender systems on users

Alistair Knott
Co-Lead
Department of Computer Science
University of Otago

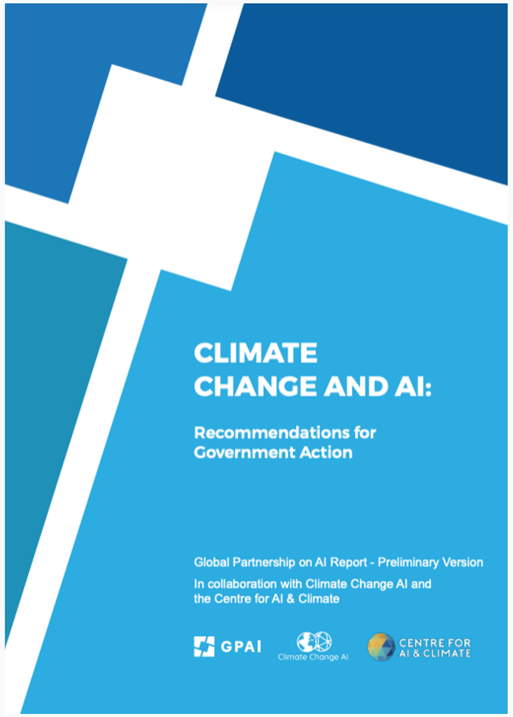
Kate Hannah
Co-Lead
Te Pūnaha Matatini
University of Auckland

Dino Pedreschi
Co-Lead
Professor of Computer Science
University of Pisa

Our new Co-Chair: 2022-24

Catherine Régis
Co-chair Responsible AI
Canada Research Chair in Health Law and Policy
Professor of Law
Université de Montréal

A Responsible AI Strategy for the Environment



CLIMATE CHANGE AND AI:

Recommendations for Government Action

Global Partnership on AI Report - Preliminary Version

In collaboration with Climate Change AI and the Centre for AI & Climate

GPAI Project *RAISE*

Responsible AI Working Group

Nicolas Miailhe, Committee Co-Lead
Raja Chatila,Committee Co-Lead
Marta Kwiatkowska, Committee Steering Group

Overall Objective

"Develop a global responsible AI adoption strategy for climate action and biodiversity"

Short-term Objectives

"Create a roadmap of AI & Climate action ahead of the COP-26"

UN CLIMATE CHANGE CONFERENCE UK 2021

- Analyse responsible AI **benefits & risks**
- Build **roadmaps** for govs, IGOs, and research
- Develop catalogue of **high-impact AI use cases**

Long-term Objectives

- Strengthen and expand the **climate action roadmap**
- Work with **institutional partners** to anchor the climate action roadmap at the **COP** and other forums

ECMWF

UNEP
International Resource Panel
ipbes
UNFCCC
ipcc INTERGOVERNMENTAL PANEL ON climate change

- Expand the scope to include **biodiversity promotion** in other GPAI projects
- Develop an **impact and risk assessment framework** harnessing AI for climate action and biodiversity preservation responsibly

GPAI PARIS SUMMIT 2021

Climate Change and AI

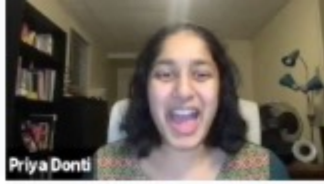
Recommendations for Government Action

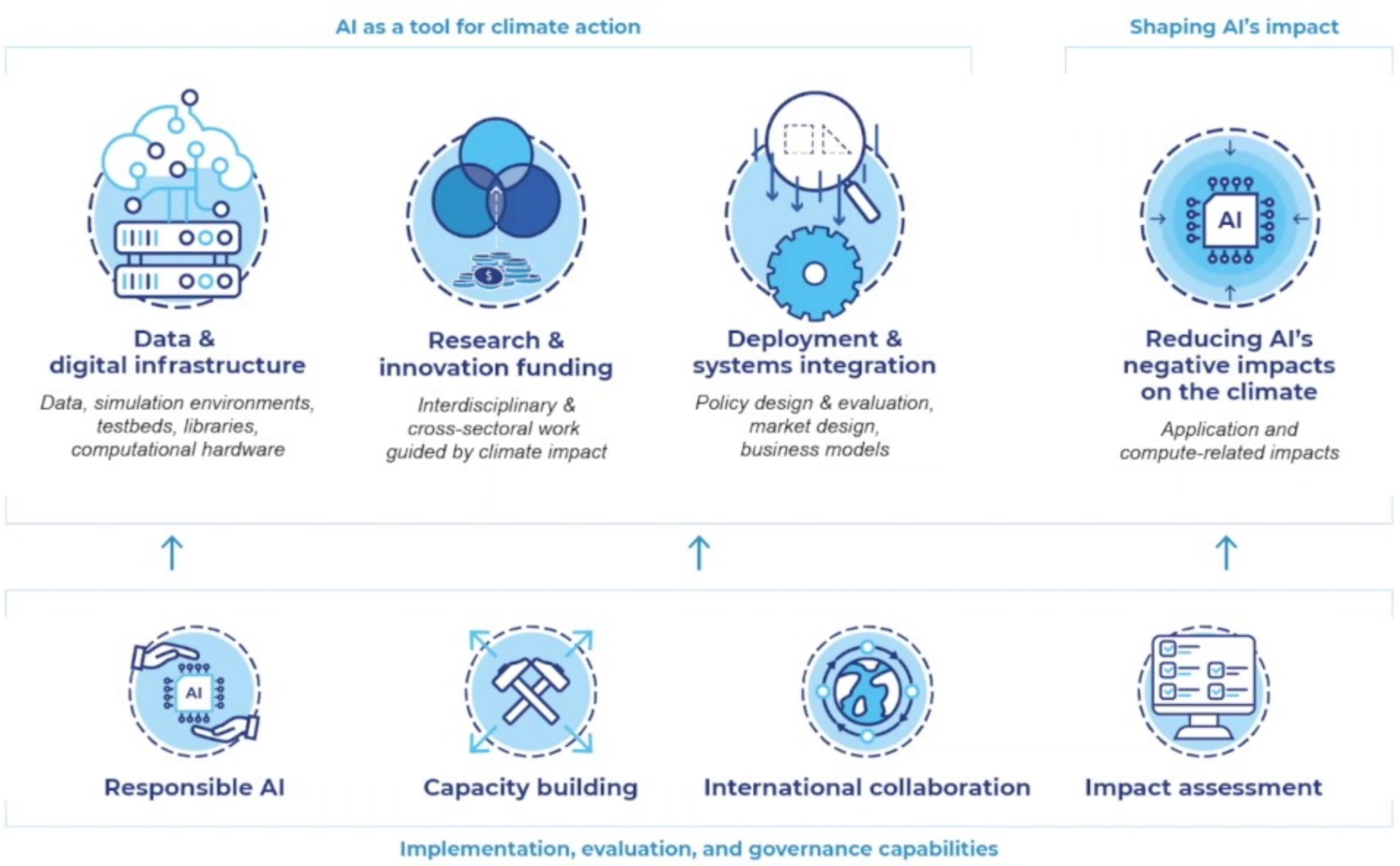
CENTRE FOR AI & CLIMATE

Climate Change AI

Report developed in collaboration between members of Climate Change AI and the Centre for AI & Climate, and experts in the Global Partnership on Artificial Intelligence’s Committee on Climate Action and Biodiversity Preservation, as part of the broader working group on Responsible AI. The report reflects the personal opinions of the authors and does not necessarily reflect the views of the experts’ organizations, GPAI, the OECD, or their respective Members.

Report Structure & Areas of Impact





AI as a tool for climate action

Data & digital infrastructure

Data, simulation environments, testbeds, libraries, computational hardware

Research & innovation funding

Interdisciplinary & cross-sectoral work guided by climate impact

Deployment & systems integration

Policy design & evaluation, market design, business models

Shaping AI's impact

Reducing AI's negative impacts on the climate

Application and compute-related impacts

Responsible AI

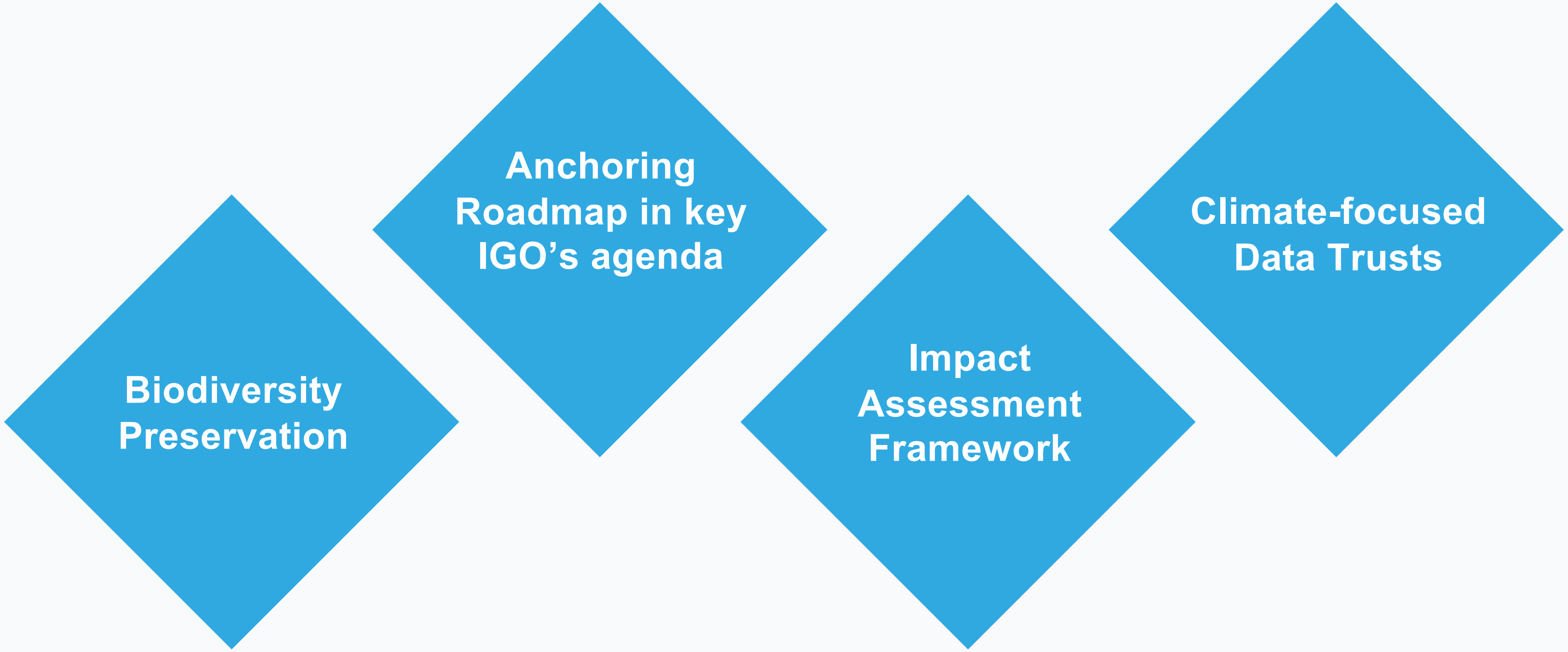
Capacity building

International collaboration

Impact assessment

Implementation, evaluation, and governance capabilities

Work Plan 2022



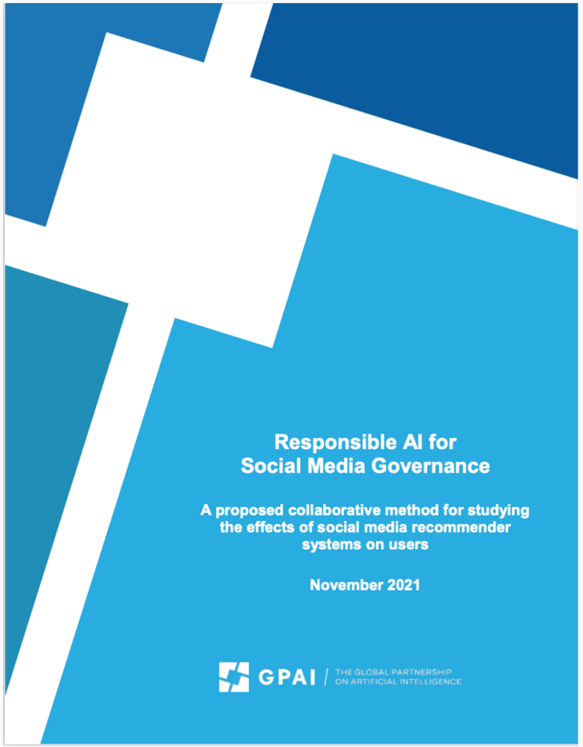
Biodiversity Preservation

Anchoring Roadmap in key IGO’s agenda

Impact Assessment Framework

Climate-focused Data Trusts

Responsible AI for Social Media Governance



Responsible AI for Social Media Governance

A proposed collaborative method for studying the effects of social media recommender systems on users

November 2021

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

Three related sub-projects

COMMUNITY CONSULTATION

Asking the citizens of a given country (NZ) more generally what counts as ‘harmful content’

TECHNICAL

Working with social media companies to investigate whether their recommender systems move users towards ‘harmful content’

LEGAL/POLICY

What’s the law/policy basis for the proposed engagement with social media companies?

Focus of technical project

Rec Sys

Focus is on the **recommender systems** that deliver content into social media users’ ‘feeds’.

Learning

RecSys are AI/ML systems that learn about what each user likes to engage with. Through learning, RecSys deliver content that’s personalised to each user.

Scientists have concerns about how Recommendation Systems learn

- Recommendation Systems learn from seeing which feed items a user clicks on (or otherwise engages with)
    - **user clicks → recsys learning**
- But the user chooses from a list of items the Rec Sys already thinks she will like
    - **recsys learning → user clicks**
- There’s a **feedback loop** here, which can lead to instabilities.
- Users also show certain systematic **biases** in their clicks:
    - A bias towards ‘moral emotions’, and negative sentiment
    - A bias towards content about political out-groups
    - A bias towards false information.
- If the Recommendation System eflects these biases, the instabilities could **lead users in harmful directions**.

Do Recommendation Systems lead users towards harmful content?

- **Prima facie concern** comes from theoretical models and simulations.
- But obviously, it must be tested on **real social networks**.
- Most studies are conducted **externally** to social media companies.
- But external methods all have limitations.
  - **Population studies** compare demographic **groups** with different Internet behaviours → *confounding variables*
  - **User behaviour studies** get data from **volunteer social media users** → *sampling problems*
  - **Robot user studies** examine the consequences of following recommended links → *robots aren't real users*
- The biggest problem: to test if a Recommendation Systems has **causal effects on users**, we must **intervene** on the Recommendation Systems—and that can only be done **inside companies**.

How companies study the effects of Recommendation Systems on users

- Social media companies are constantly trying out different versions of their Rec Sys on users, and picking the ones which are ‘best’, by their criteria.
- They use many criteria, but centrally they are looking for Rec Sys that maximise user engagement with their platform.
- Company-internal methods avoid the problems of external methods:
    - No confounding variables.
    - No sampling problems.
    - Studies are of real users, on real social media platforms.
    - Studies test proper causal hypotheses about Rec Sys effects.

A proposed ‘fact-finding study’

- We propose a method for a government to work with a company, to ask whether its Rec Sys are moving users towards ‘harmful content’.
    - We aim to trial this method in **New Zealand**, as a case study.
    - We are focussing on **‘Terrorist and Violent Extremist Content’** (TVEC), to fit in with this year’s **Christchurch Call** workstream.
- Our fact-finding study augments **company’s existing studies** of Rec Sys effects on users, with new metrics, that **measure users’ engagement with ‘harmful material’**.
    - Our focus is on metrics that gauge users’ relationship towards TVEC.
    - The study is to be co-designed by the company and a group of independent experts.

The proposed ‘fact-finding study’ will ask two questions

1) Do different Recommendation Systems have different effects on users’ relation towards harmful content?

2) Do Recommendation Systems that ‘maximise user engagement’ also drive users towards harmful content?

Frances Haugen’s recent revelations suggest the answer for Facebook’s RecSys may be ‘yes’ in both cases.

But we can’t rely on one-off disclosures based on unseen documents! We need a way of surfacing scientific findings about Rec Sys effects.

Publishing the results of the fact-finding study

Our proposal is that the **results of a fact-finding study** requested by a government are **published in a scientific paper**.

Our method is **safe**:
doesn’t compromise **company IP**

- delivers transparency about the effects of Recommendation Systems, not their internal workings

doesn’t compromise **user privacy**

- measures of user behaviour are aggregated over large user groups

Twitter has recently published a paper describing exactly the kind of fact-finding study we envisage

- Huszár et al. “*Algorithmic Amplification of Politics on Twitter*”, posted 21 October 2021

The community consultation project:

Community Consultative Processes for Grounded, Situated Harm-focused Responses

Question 1

How do the **NZ communities** who experience the most harm online from hateful expression, dangerous speech, and misinformation **define those harms** and their lived experiences of them?

Question 2

How can mediation, moderation, regulation, and categorisation as **co-developed and co-utilised tools** mitigate against those harms and improve communities' experiences of online spaces?

Key method: meetings (hui) with community groups

Current online harm/disinformation in Aotearoa remains at high threat level

Researchers will be using word-of-mouth, trust-based channels for inclusion of participants in hui, to keep people safe.

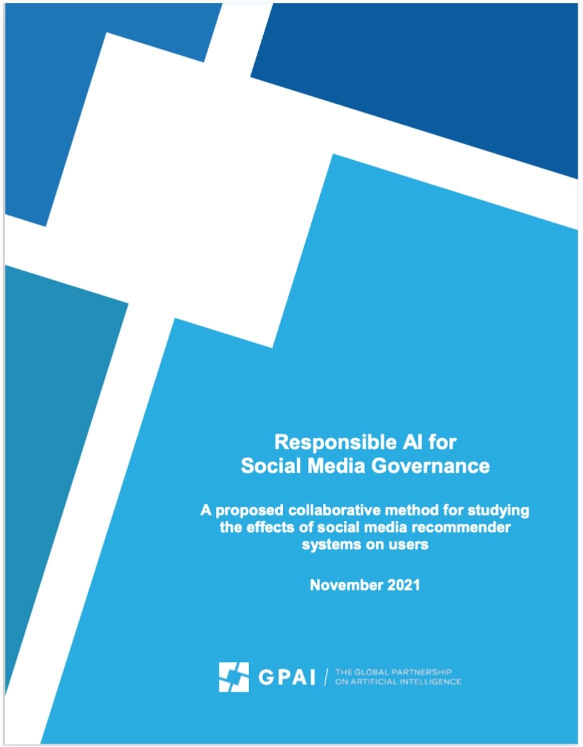
Community meeting (hui)

Initial hui in June 2021. Online community hui (zui) in Sept Oct. 2021. Preparation for larger hui's in 2022, focus on classification and categorisation for Aotearoa.

The law/policy project

investigates the legal/policy basis for the proposed fact-finding exercise.

- For details, please see the written report!



Responsible AI for
Social Media Governance

A proposed collaborative method for studying the effects of social media recommender systems on users

November 2021

GPAI / THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

GPAI / THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

THANK YOU

Contact:
info@ceimia.org