How can we design ethical AI systems:

Connecting ethics to policy and design

Shalaleh Rismani
Chief Innovation Officer, Generation R Consulting
Director, Open Roboethics Institute

GENERATION R

ORI
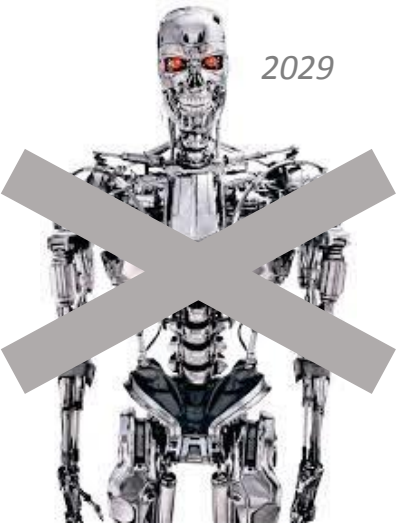Open Roboethics Institute

Meet the team

AJung Moon

Jason Miller

Shalaleh Rismani

What do we mean by AI?

ARTIFICIAL GENERAL INTELLIGENCE

2029

SING~~ULARITY~~

2017

2015

ARTIFICIAL INTELLIGENCE

BIG DATA

MACHINE LEARNING

DATA SCIENCE/ DATA ANALYTICS

DEEP LEARNING

pepper

GENERATION R

4

THE AUTONOMY SPECTRUM

MACHINE
AUTONOMY

**Today's Technology &
Human Decisions**

HUMAN
AUTONOMY

GENERATION R

What do we mean by "*Ethical Design*"?

A Trend

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

General Principles (from Version 2)

1. Human rights

2. Well-being

3. Accountability

4. Transparency

5. Awareness of misuse

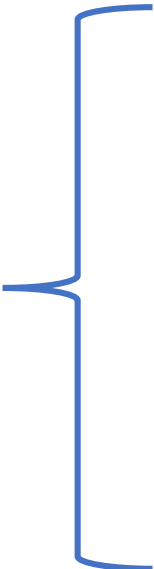# UK Data Ethics Framework for Public Sector

Key Principles

1. Start with a clear user need

2. Use the data that you need to have

3. Make your work transparent

4. Embed data use responsibly

5. Understand limitation of data

6. Use robust practices

7. Be aware of legislations and codes of practice

Technology Companies have developed their own AI ethics principles in 2018.

- Microsoft
- Google
- SAP

Driven by human values

Avoid reinforcing negative bias

Privacy by design

Strive for high level of transparency

Consider those whom are accountable

What does this mean for businesses implementing AI, *particularly recommender systems*?

GENERATION R

A Case Study Organization:
Technical Safety BC

Use Case:
An algorithm that predicts likelihood of hazard and directs inspectors

GENERATION R

Technical Safety BC in 2017

Are we developing our AI system with our company values mind?

What is our bounding box for design of technology when it comes to ethics?

Is our technology meeting high ethical standards?

What will this meant to our employees?

What will this mean to public and public safety?

AI ETHICS RISK

Probability of ethical harm occurring from the frequency and severity of exposure to a hazard

{ **societal** issues

acceptance issues

public **perception** issues

failure to launch

commercial/financial loss

deterioration of **work culture**/environment

environmental issues

... }

ETHICS ASSESSMENT

Understanding of ethical risks associated with a project with the goal of creating a roadmap that helps manage the risks

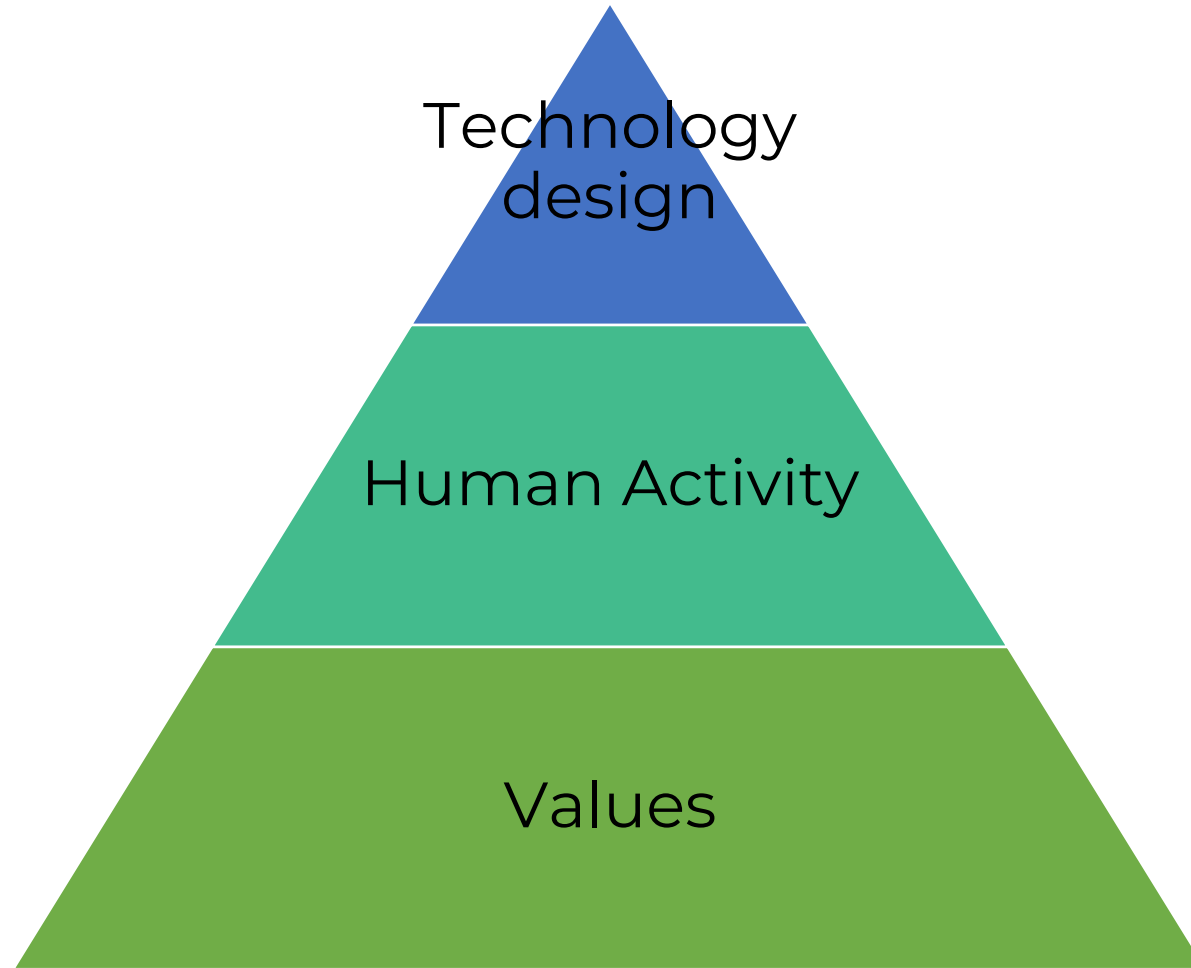{ **transparency** and interpretability;

discrimination and **fairness**;

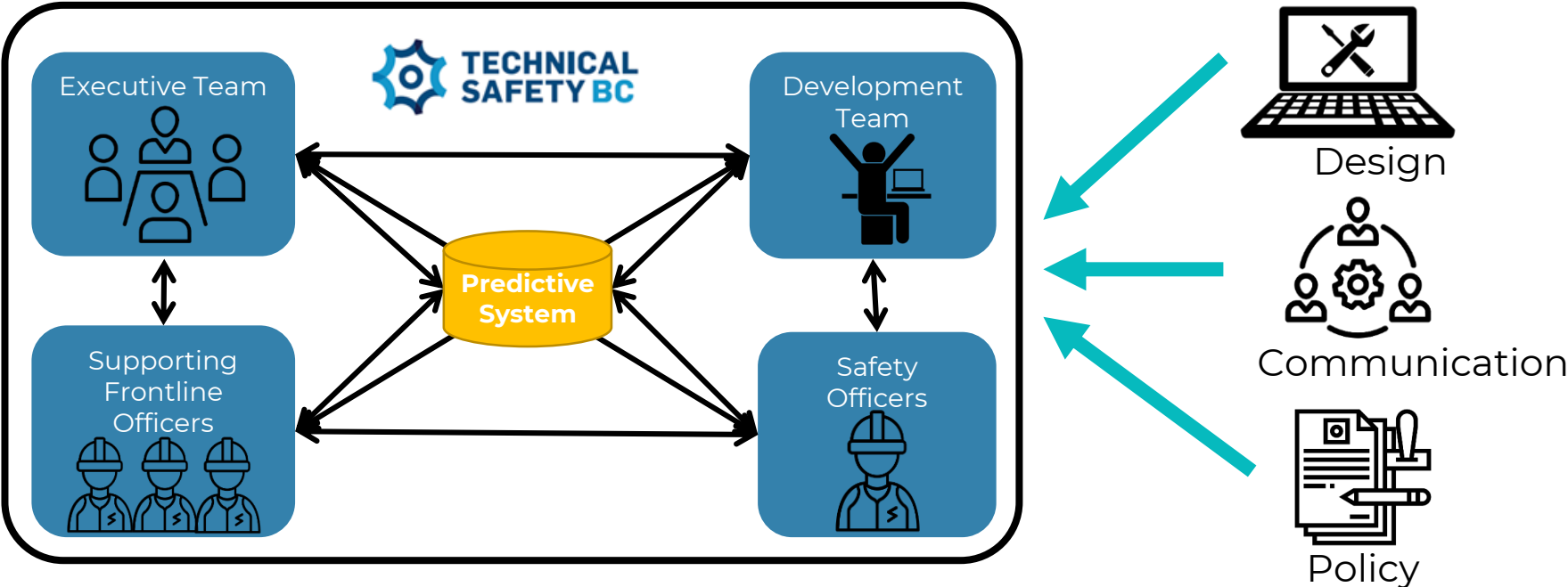stakeholder perception, awareness, and **trust**;

privacy and security;

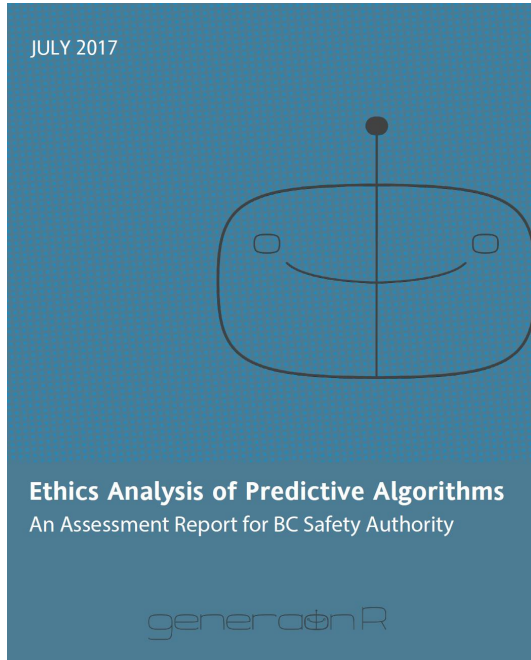responsibility and **accountability**;

impact on **autonomy** and jobs }

Technology design

Human Activity

Values

AN AI ETHICS ROADMAP

(Value conflicts -> Value alignment) -> **a Roadmap**

Ethics Analysis of Predictive Algorithms
An Assessment Report for BC Safety Authority

generation R

Full Report
Free & Available Online:

http://bit.ly/AlgEthics

Transparency and Interpretability

Autonomy

Public Perception

Fairness

Accountability

GENERATION R

Analysis Results:
Autonomy and Impact on Jobs

Making a decision about what to inspect

Will this affect my job security?

Predictive System

Leveraging local network

Using knowledge and expertise
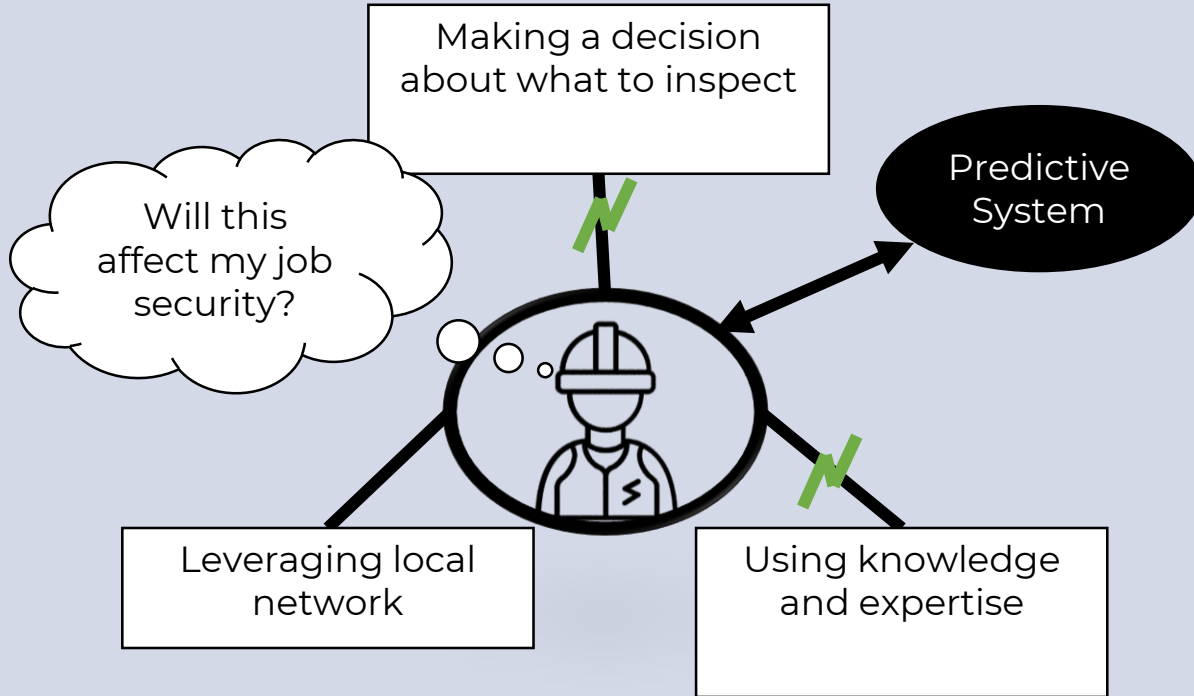
Key elements of autonomy for safety officers

- Presentation of Output
- Feedback Mechanism

- Foster Discussions

GENERATION R

Analysis Results:
Fairness & Discrimination

Predictive System

Input → [computer/model] → Output
P(1)= 40%

Data entry and interpretation
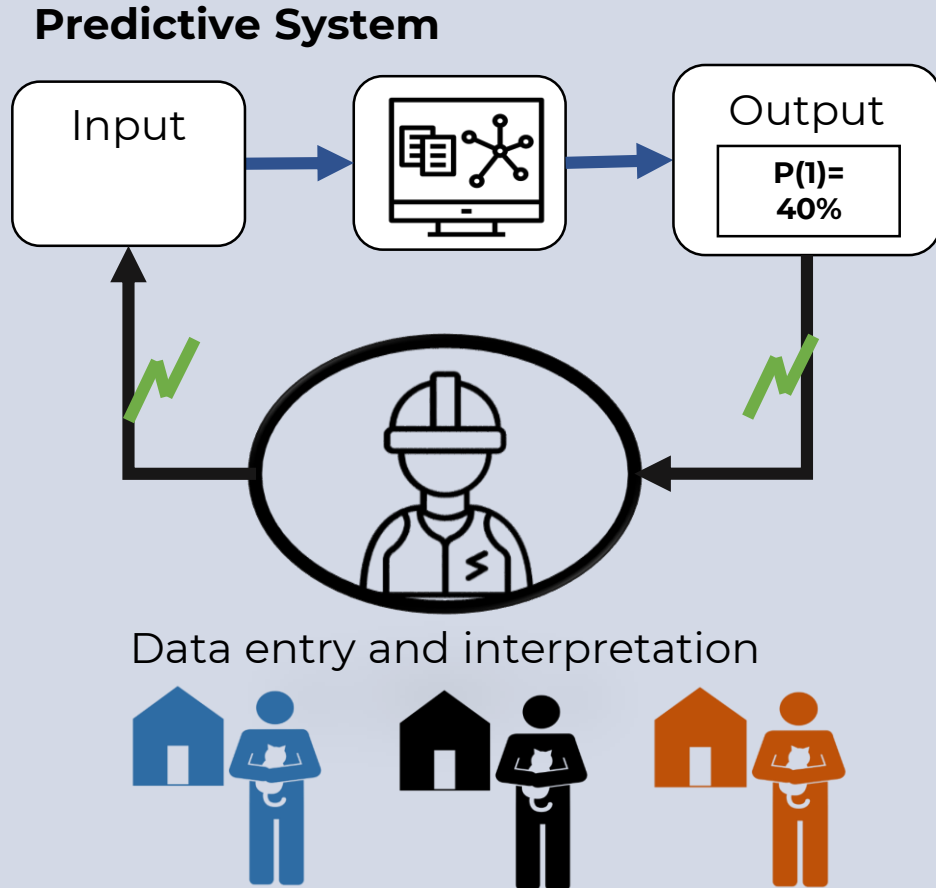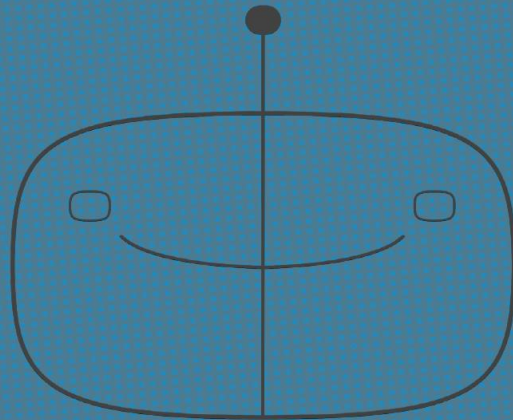
- Foster consistent data entry
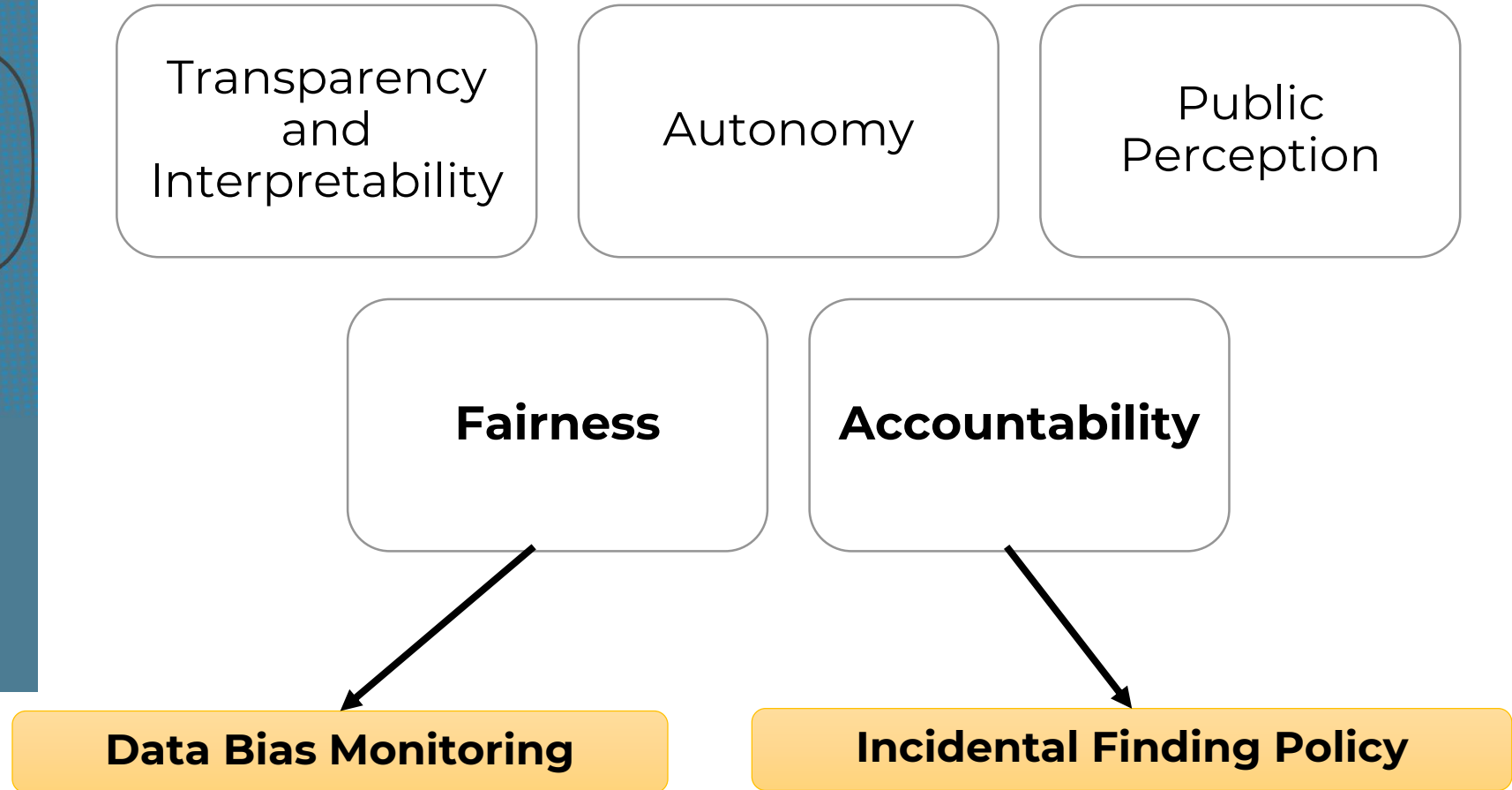- Monitor for discriminatory patterns

- Define fairness within the organization
- Means to monitor discriminatory patterns

GENERATION R

Ethics Analysis of Predictive Algorithms
An Assessment Report for BC Safety Authority

Taking it one step further

Transparency and Interpretability

Autonomy

Public Perception

Fairness
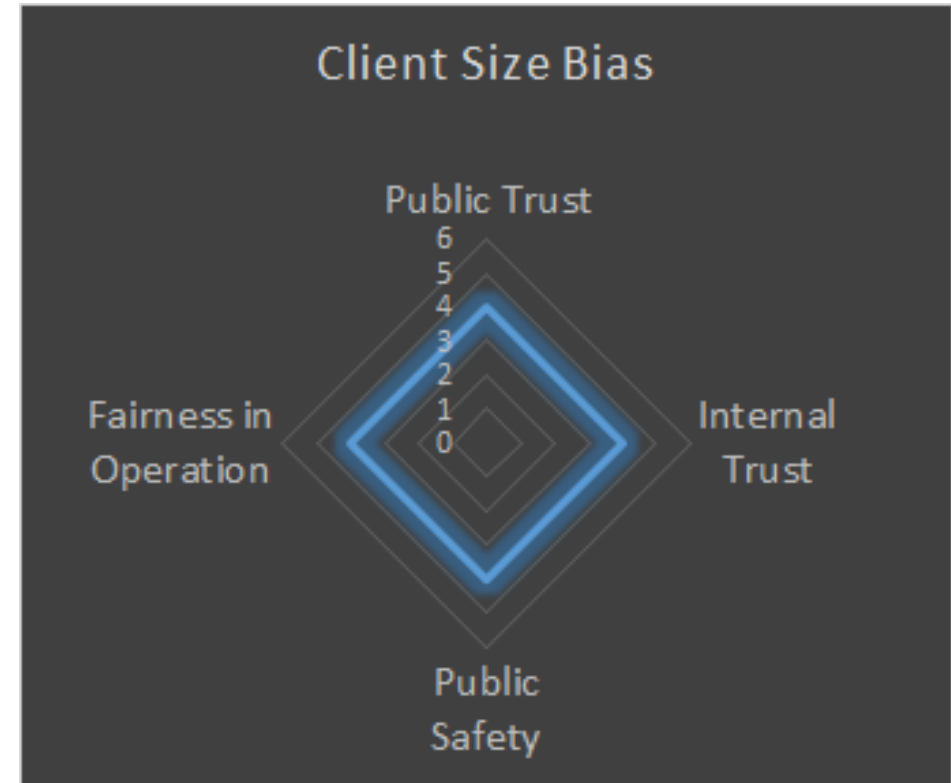
Accountability

Data Bias Monitoring

Incidental Finding Policy
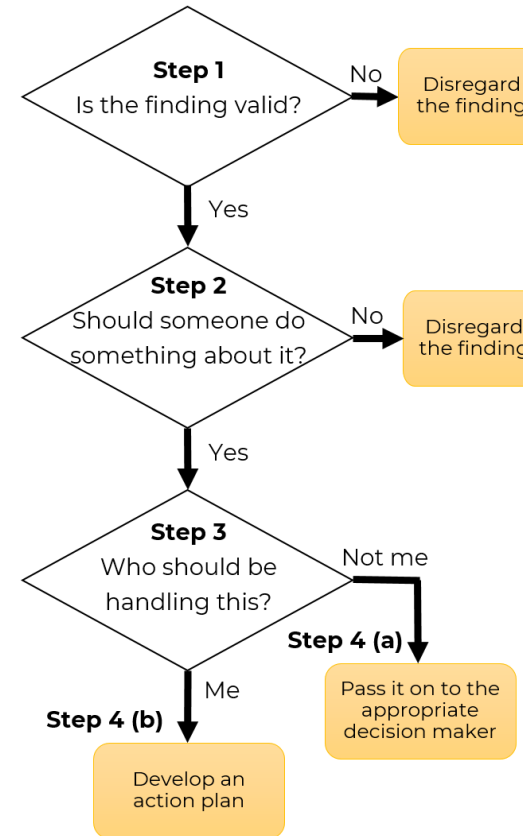
GENERATION R

Bias Monitoring based on Input Data Set

1. Public Trust
2. Internal Trust
3. Public Safety
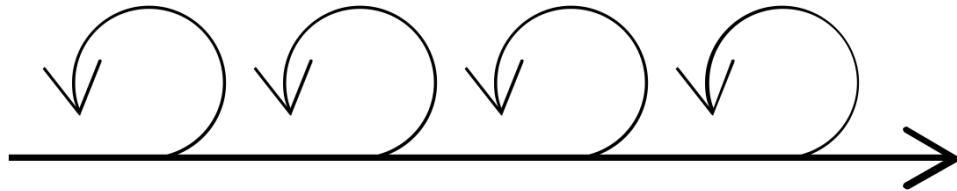4. Fairness in Operation

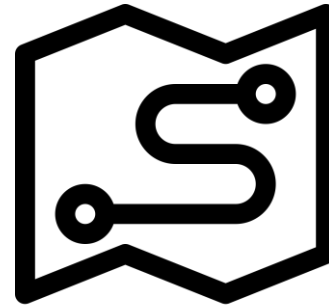GENERATION R

Incidental Finding Policy

- Conducted an organizational value workshop

- Co-developed a policy to handle unintended finding

- Enable everyone to take an action when they come across an incidental finding

Step 1
Is the finding valid? — No → Disregard the finding
↓ Yes

Step 2
Should someone do something about it? — No → Disregard the finding
↓ Yes

Step 3
Who should be handling this? — Not me → **Step 4 (a)** Pass it on to the appropriate decision maker
↓ Me
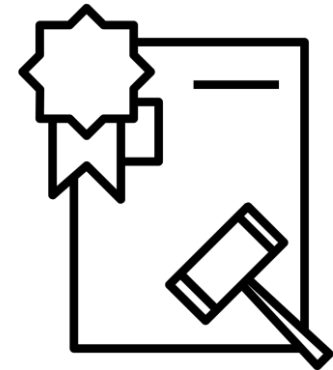**Step 4 (b)** Develop an action plan

GENERATION R

What is next?

Iterate as Purpose Changes

vs.

Ethics Roadmap vs. Legal Action

Key takeaways

- Ethical principles for AI exist – by various organizations, companies and governments
  - However, we do not have an international set of guidelines yet!
- The challenge is to

  1. Understand the ethical needs of each organization and AI application through a systematic process

  2. Identify the particular solutions for these needs

What particular ethical
challenges does your AI
system (i.e. recommender
system) have?

GENERATION R

Thank you!

GENERATION R

ORI
Open Roboethics Institute