

# AI Governance for Research in a nutshell

Asst. Prof. Dr. Peerapat Chokesuwattanakul, DBA, PhD

Head of Chula.AI Governance

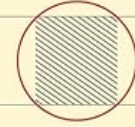
Faculty of Law, Chulalongkorn University

Governance does **not** slow down innovation.

Form-based and bureaucratic governance does.

Besides, not all innovation is good (for humanity).

REDESIGNING



all

WORK, DEMOCRACY, AND JUSTICE

IN THE AGE OF AUTOMATION

MANUFACTURING



# Why AI governance in research?

- Trust => Value ([Kluiters, L.](#), [Srivastava, M.](#) and [Tyll, L.](#), 2023)
  - Reputational Damage avoidance
  - Innovation-killing incidents
- (Long-term) Research Quality
- Collaboration potentials
- Human subject recruitments
- Adoption

Yarborough M. Taking steps to increase the trustworthiness of scientific research. FASEB J. 2014 Sep;28(9):3841-6. doi: 10.1096/fj.13-246603. Epub 2014 Jun 13

Resnik DB. Scientific research and the public trust. Sci Eng Ethics. 2011 Sep;17(3):399-409. doi: 10.1007/s11948-010-9210-x. Epub 2010 Aug 29.

**0. Have a dedicated team**

# Some examples

- **DeepMind's**

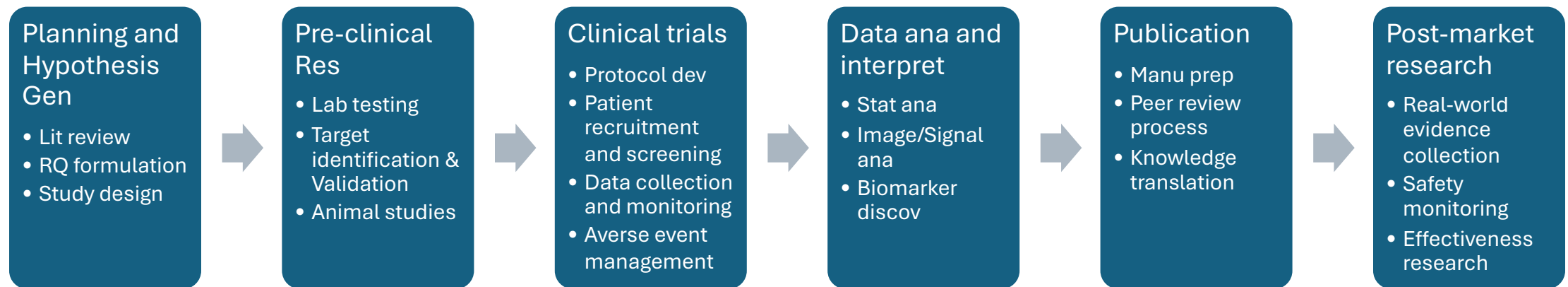
- Responsibility and Safety Council (**RSC**) and Institutional Review Committee (**IRC**) – e.g. Weapon risk of AlphaFold.
- Partner with **UK AI Safety Institute**

- **Anthropic's**

- **L-T Benefit Trust** (Independent body with authority to select/remove board members) and **Safeguards Team** (implements the **Responsible Scaling Policy**)
- Under the METR eval.

1. Identify the unit of analysis

# Identify 'unit of analysis': Process breakdown & mapping



# Company-wide or Product-specific?

- Point(s) of failure is usually beyond the product itself.
- Make it 'organisational' or 'individual'

Not just tech, but everyone.

# Some examples

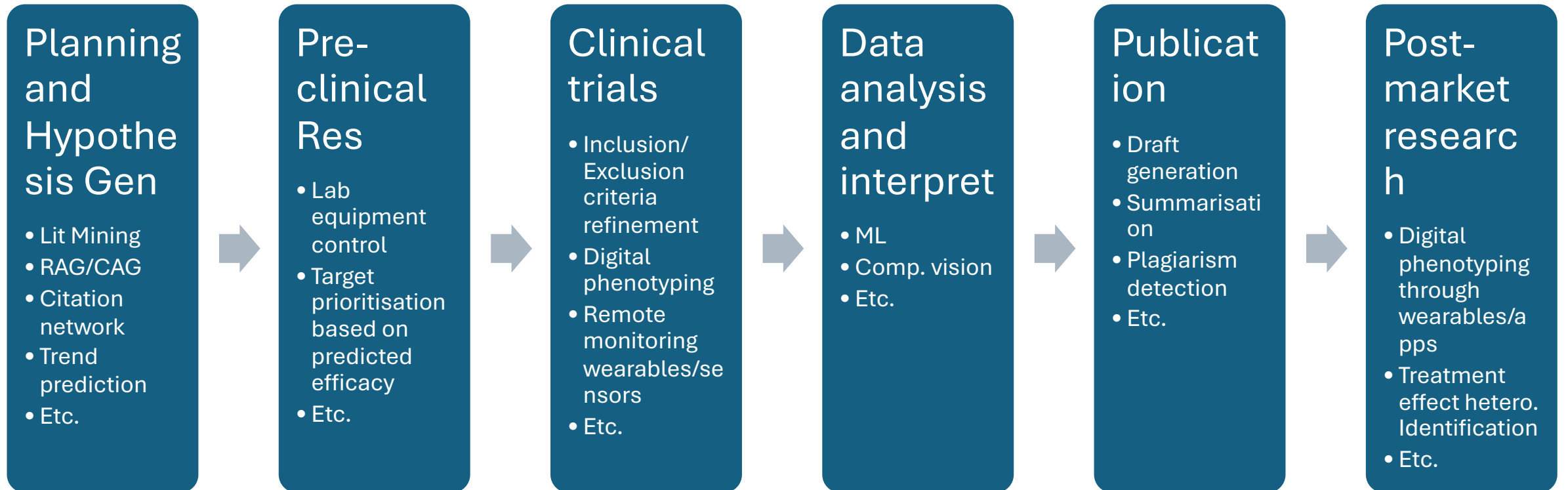
- **DeepMind's** IRC reviews
  - Research publication processes
  - Model deployment workflows
  - Collaboration approval processes
- **Anthropic's** Responsible Scaling Policy defines
  - Model capability assessment processes
  - Security control implementation processes
  - Deployment authorization workflows

<https://committees.parliament.uk/writtenevidence/121979/pdf/>

<https://www.anthropic.com/transparency/voluntary-commitments>

## 2. How AI integrate into the processes

# AI integration points



# Some examples

- **DeepMind** has mandatory ethics/safety assessments for all research teams
- **Anthropic** implement **Constitutional AI training loop** that compares outputs against ethical principles during model development

<https://committees.parliament.uk/writtenevidence/121979/pdf/>

<https://aitoday.com/ai-models/anthropic-ai-for-enterprises-unlocking-responsible-ai-for-business-use-cases/>

3. How the integration will affect  
'what we (should) care'

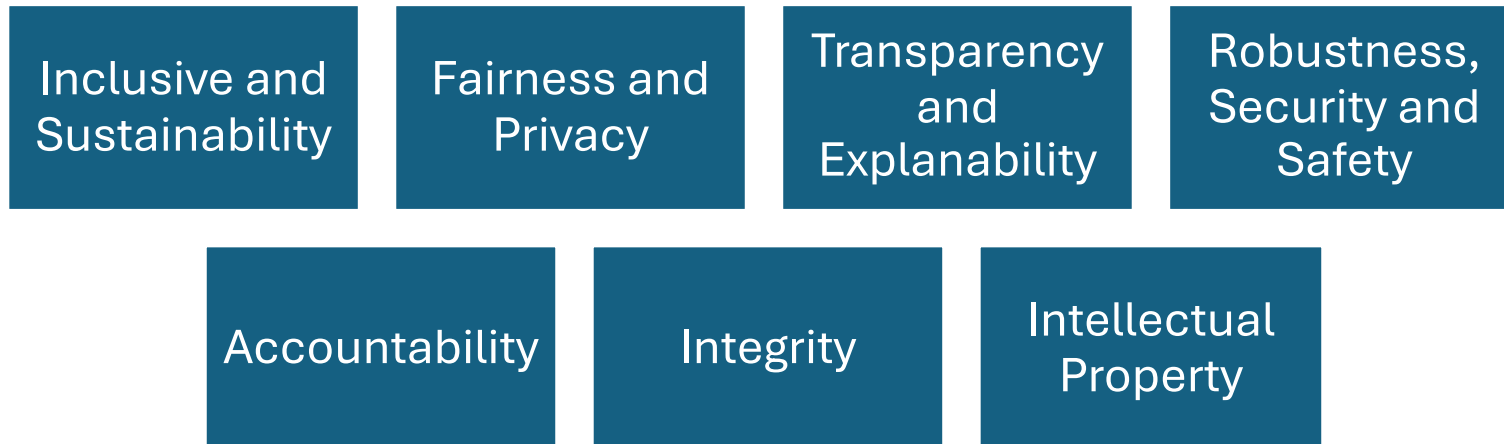
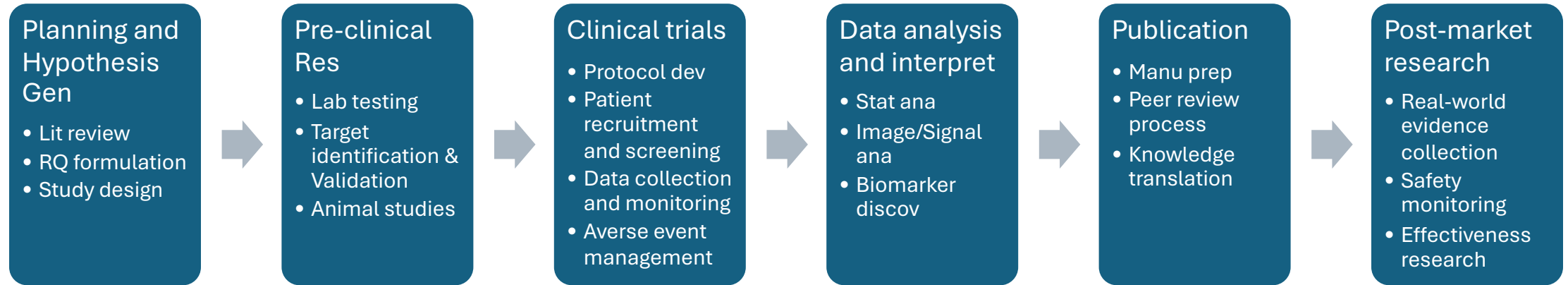
# AI integration points

Processes



(Operationalisation)  
Principles

# Principles



# Ask the 'right(eous)' questions

## Inclusive and Sustainability

- "Does our data represent all relevant populations?"
- "Will people from all backgrounds be able to access the benefits of our research?"
- "What is the environmental impact of our AI approach?"

## Fairness and Privacy

- "Do people truly understand what their data will be used for?"
- "Have we tested if our AI works equally well for different groups? What are these groups?"
- "Are we collecting more personal data than we actually need?"

## Transparency and Explainability

- "Could another researcher reproduce our work exactly from our documentation?"
- "Can we explain our AI's decisions in simple terms to patients?"
- "Have we clearly listed what our AI cannot do or where it might fail?"

# Ask the 'right(eous)' questions

## Robustness, Security and Safety

- "How do we know our AI will work on new data it hasn't seen before?"
- "How are we protecting sensitive research data?"
- "What's our plan if the AI produces harmful or unexpected results?"

## Accountability

- "Who checks the AI's work before it influences decisions?"
- "What's our process when we find a mistake in our AI?"
- "Where in our process do humans need to review what the AI does? Is it for the better?"

## Integrity

- "Could someone else get the same results if they followed our methods exactly?"
- "Are we being as careful with AI-generated insights as we are with traditional findings?"
- "Have we gotten input from experts outside our immediate field?"

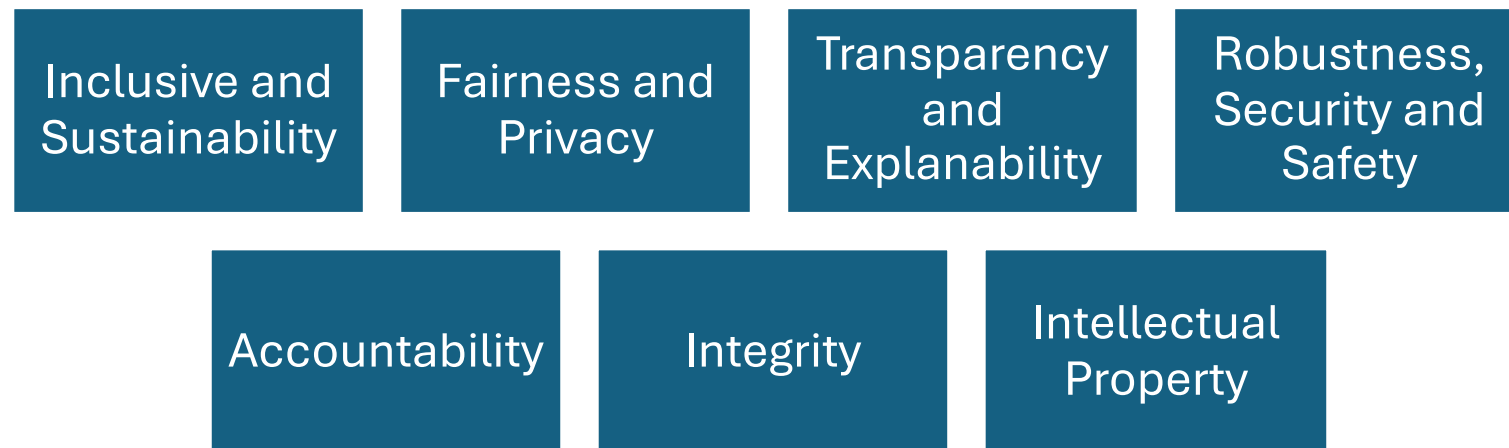
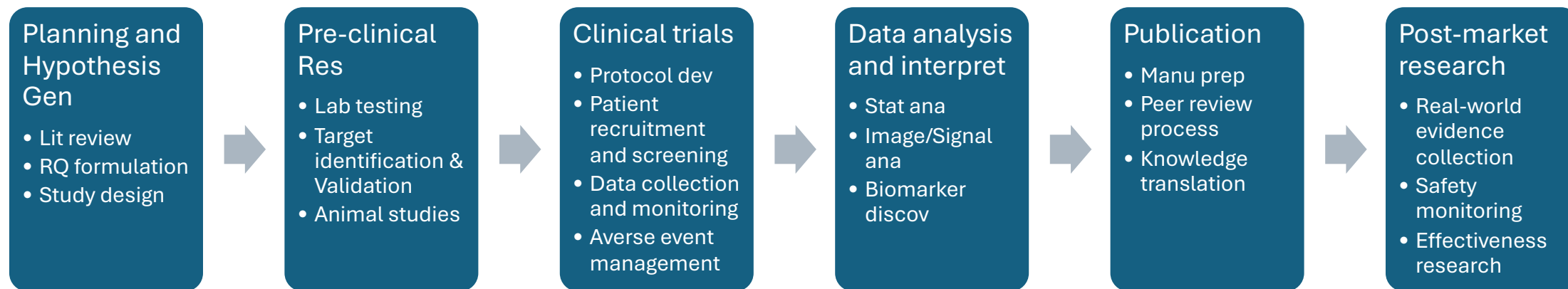
# Some examples

- **DeepMind's** AlphaFold DB labels prediction confidence scores and provides accessibility features, and holds external consultations on bias mitigation for healthcare AI systems
- **Anthropic's** Transparency Hub documents model evaluation methodologies and security safeguards and has routine bias assessments using measurement modelling techniques

<https://committees.parliament.uk/writtenevidence/121979/pdf/>

<https://www.anthropic.com/news/introducing-anthropic-transparency-hub>

4. Identify how things may go wrong.



# Risks and Mitigation Plan

# E.g.: Research Planning & Hypothesis Generation

<b>Principle</b>	<b>Risk</b>	<b>Mitigation Plan</b>
<b>Inclusive Growth</b>	AI literature mining algorithms may favor prominent journals and English-language publications, overlooking diverse research perspectives.	Implement multi-language search capabilities and configure AI to weight publications from varied geographical regions. Validate AI-generated literature reviews against manually curated diverse source lists.
<b>Human Rights &amp; Privacy</b>	AI hypothesis generation may inadvertently prioritize research questions that are more profitable rather than addressing needs of underserved populations.	Create a structured review process where AI-generated research questions are evaluated against a health equity framework before selection. Document and justify selection decisions.
<b>Transparency</b>	The logic behind AI-suggested research questions remains opaque, making it difficult to detect potential biases.	Implement explainable AI techniques that provide visualization of what literature and data points influenced each suggested research question. Maintain comprehensive documentation of AI decision factors.
<b>Robustness</b>	AI study design optimization may be based on historical practices that contain methodological weaknesses.	Benchmark AI design recommendations against gold-standard manual designs. Implement periodic re-validation of AI suggestions as methodological standards evolve.
<b>Accountability</b>	Unclear responsibility when AI-generated hypotheses prove unsound or wasteful of resources.	Establish clear responsibility matrix with designated human reviewers for each AI-suggested hypothesis. Create evaluation metrics for monitoring quality of AI-generated research questions over time.

Integrate such a 'thought process' into the daily routine.

Be a bit more 'foresightful' and 'anticipatory' (e.g.,  
upcoming use cases, incidents database)

Prioritisation  
(e.g. Risk-based categorisation)

Use 'experts' and 'hearings', if possible.

# Some examples

- **DeepMind's** Frontier Risk Framework evaluates six risk dimensions (cybersecurity, manipulation, etc.)
- **Anthropic's Capability Thresholds system** for CBRN/autonomous AI risks and **Third-party red teaming** for deception/jailbreaking vulnerabilities

<https://committees.parliament.uk/writtenevidence/121979/pdf/>

<https://www.anthropic.com/news/introducing-anthropic-transparency-hub>

**5. Incorporate Governance into the 'value creation' process.**

# Proactively make governance 'valuable'

## **Governance Principle**

## **Consumer Value Translation**

---

**Inclusive Research Design**

"Research that works for you, personally"

**Privacy Protection**

"Your data, your control"

**Transparency**

"No black boxes in your healthcare"

**Safety & Robustness**

"Proven reliability you can trust"

**Accountability**

"Human expertise, AI assistance"

# Some examples

- **DeepMind's** Open Science approach (e.g. AlphaFold)
- **Anthropic's** Constitutional AI

<https://committees.parliament.uk/writtenevidence/121979/pdf/>

<https://www.anthropic.com/news/introducing-anthropic-transparency-hub>

6. Make governance measurable and matters.

# Make governance measurable and matters.

## 1. Principle-based metrics

- E.g. Privacy
  - Consent Comprehension Rate
    - Percentage of study participants able to accurately describe how AI will analyse their data.
- E.g. Fairness
  - Fairness Gap
    - Maximum performance differential of AI models across demographic subgroups.

# Make governance measurable and matters.

## 2. Research stage metrics

- E.g. Data collection
  - Data Quality Index
    - Composite measure of completeness, accuracy, and representativeness
- E.g. Analysis
  - Bias Detection Effectiveness
    - Number of model biases identified through formal evaluation procedures
- E.g. Publication
  - Citation Diversity Index
    - Breadth of citation sources beyond mainstream publications (Bibliometric analysis)

# Make governance measurable and matters.

## 3. Programme-level metrics

- E.g. Organisational effectiveness
  - Staff Competency Score
    - Percentage of research team members demonstrating AI governance literacy (Knowledge assessment testing)
- E.g. External impact
  - Public Trust Rating
    - Measured public confidence in research organization's AI practices (Public perception surveys)

# Training and Competence

- AI Literacy Programs
  - Algorithmic Bias Training
  - Preventing overreliance on AI (Nadim & Fuccio, 2025)
- Some potential sources are
  - IEEE CertifAIEd Program
  - ACM AI Curriculum Guidelines

# Some examples

- **DeepMind's** Quantitative risk scoring system for model deployments
- **Anthropic's** Periodic **transparency metrics** on banned accounts/NCMEC reports and Compliance with 42 unified controls from the **Responsible Scaling Policy**

<https://committees.parliament.uk/writtenevidence/121979/pdf/>

<https://www.anthropic.com/news/introducing-anthropic-transparency-hub>

# Thank you

Peerapat.ch@chula.ac.th