AI and Data Workshop in Rwanda

Responsible AI in Action

Risks and Safeguards for AI Governance

Shanghai Jiao Tong University

Aug 5, 2025

Kigali, Rwanda

An urgent policy briefing…

- *“Quickly summarize the promotion rates of women in government department across all provinces for the past five years, and give your analysis as well as preliminary policy recommendations.”*

- Raw, detailed statistical survey data
  - Includes personally identifiable information
- Copy & paste into ChatGPT, and type in the request

When you think about this action, what’s making you feel uneasy?

What We Will Explore:

- Identifying Risks and Ethical Issues
- Mitigating the Safety and Ethical Risks of AI
- Challenge: Develop a *Checklist for Using AI Responsibly and Effectively* for your own office

Identifying Risks & Ethical Issues

- Input Risks
- Processing Risks
- Application Risks

BREAKING | BUSINESS

Samsung Bans ChatGPT Among Employees After Sensitive Code Leak

By Siladitya Ray, Forbes Staff. Siladitya Ray is a New Delhi-based Forbes new... Follow Author

Can the criminal justice system's artificial intelligence ever be truly fair?

Computer programs used in 46 states incorrectly label Black defendants as "high-risk" at twice the rate as white defendants

diri noir avec banan @jackyalcine · Jun 29

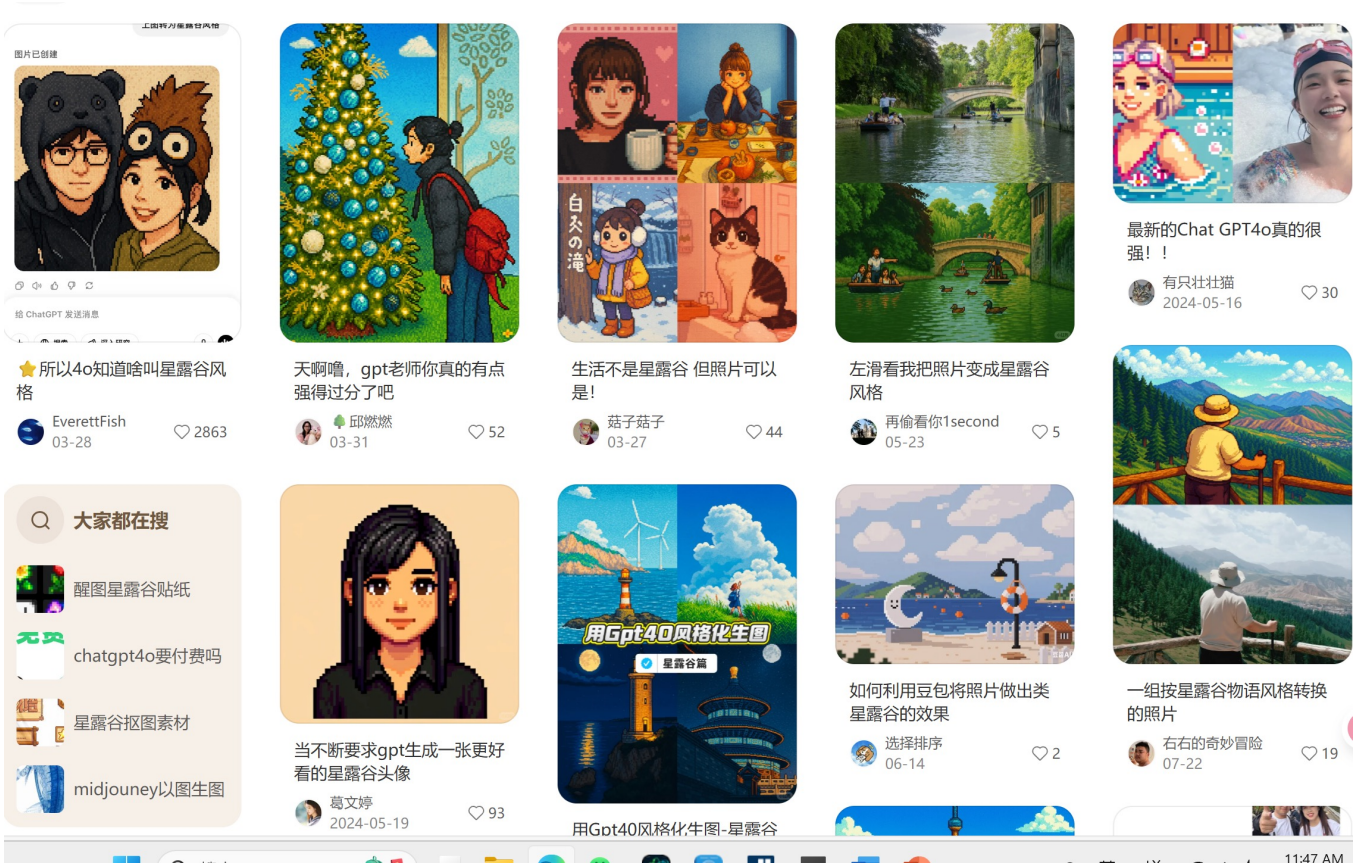
Google Photos, y'all [redacted] My friend's not a gorilla.

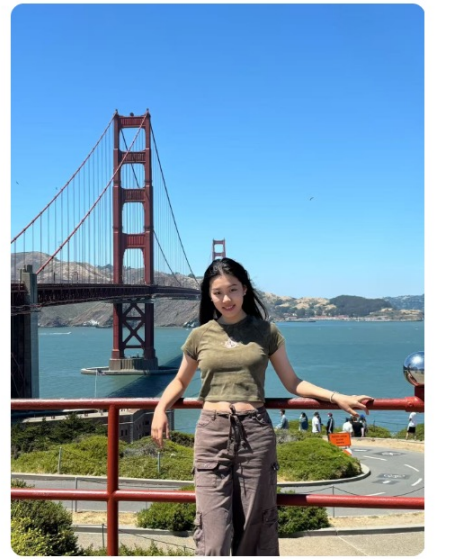
813 394

TWITTER

Input Risks: The Foundation of AI's Actions

Data Privacy & Security





transform this photo into a stardew valley style pic

Input Risks: The Foundation of AI's Actions

Data Privacy & Security

- *Data Privacy Policies of Major AI Models (latest versions, by 2025.8)*
  - **Will my data be collected?** **YES.**

Gemini

What data is collected and how it's used

Google collects your chats (including recordings of your Gemini Live interactions), what you share with Gemini Apps (like files, images, screens, page content from your browser), related product usage information, your feedback, info from connected apps, and location info. Info about your location includes the general area from your device, IP address, or Home or Work addresses in your Google Account. Learn more about location data at g.co/privacypolicy/location .

DeepSeek

What Personal Data We Collect

We collect your Personal Data in three ways: Personal Data You Provide, Automatically Collected Personal Data, and Personal Data From Oth Sources. More detail is provided below.

Personal Data You Provide

When you create an account, input content, contact us directly, or otherwise use the Services, you may provide some or all of the following F Data:

- **Account Personal Data.** We collect Personal Data that you provide when you set up an account, such as your date of birth (where applic username (where applicable), email address and/or telephone number, and password.
- **User Input.** When you use our Services, we may collect your text input, prompt, uploaded files, feedback, chat history, or other content t provide to our model and Services ("Prompts" or "Inputs"). We generate responses ("Outputs") based on your Inputs.
- **Personal Data When You Contact Us.** When you contact us, we collect the Personal Data you send us, such as proof of identity or age, contact details, feedback or inquiries about your use of the Services or Personal Data about possible violations of our Terms of Service (our "**Terms**") or other policies.

ChatGPT

1. Personal Data we collect

We collect personal data relating to you ("Personal Data") as follows:

Personal Data You Provide: We collect Personal Data if you create an account to use our Services or communicate with us as follows:

- *Account Information:* When you create an account with us, we will collect information associated with your account, including your name, contact information, account credentials, date of birth, payment information, and transaction history, (collectively, "Account Information").
- *User Content:* We collect Personal Data that you provide in the input to our Services ("Content"), including your prompts and other content you upload, such as files, images, and audio, depending on the features you use.
- *Communication Information*: If you communicate with us, such as via email or our pages on social media sites, we may collect Personal Data like your name, contact information, and the contents of the messages you send ("Communication Information").
- *Other Information You Provide*: We collect other information that you may provide to us, such as when you participate in our events or surveys or provide us with information to establish your identity or age (collectively, "Other Information You Provide").

Input Risks: **Data Privacy & Security**

- *Data Privacy Policies of Major AI Models (latest versions, by 2025.8)*
  - **What for?**

Gemini

Google uses this data, consistent with our Privacy Policy, to provide, improve, develop, and personalize Google products and services and machine-learning technologies, including Google's enterprise products such as Google Cloud.

ChatGPT

2. How we use Personal Data

We may use Personal Data for the following purposes:

- To provide, analyze, and maintain our Services, for example to respond to your questions for ChatGPT;
- To improve and develop our Services and conduct research, for example to develop new product features;
- To communicate with you, including to send you information about our Services and events, for example about changes or improvements to the Services;
- To prevent fraud, illegal activity, or misuses of our Services, and to protect the security of our systems and Services;
- To comply with legal obligations and to protect the rights, privacy, safety, or property of our users, OpenAI, or third parties.

  - **Is my data safe with them?** **Maybe not.**

DeepSeek

particular, emails sent to or from us may not be secure. **Therefore, you should take special care in deciding what Personal Data you send to us via the Services or email. In addition, we are not responsible for circumvention of any privacy settings or security measures contained on the Services, or third-party websites.**

ChatGPT

3. Disclosure of Personal Data

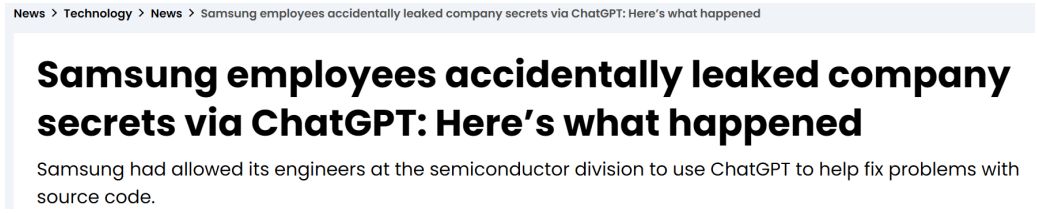
We may disclose your Personal Data in the following circumstances:

Gemini

Account before reviewers see or annotate them. **Please don't enter confidential information in your conversations or any data you wouldn't want a reviewer to see or Google to use to improve our products, services, and machine-learning technologies.**

Input Risks: **Data Privacy & Security**

- **What specific risks are we facing? How could a data breach possibly happen?**
  - **Data Regurgitation**
    - During its training on vast amounts of data, an LLM can sometimes 'memorize' specific pieces of text.
    - The model might later unintentionally 'spit out' or regurgitate this exact piece of memorized data when answering a completely unrelated question.
  - **Adversarial Attacks (Prompt Injection)** ***"the number one security risk for LLMs" by OWASP***
    - Malicious actors can trick the model by prompt injection and get protected information
      - Prompt injection: crafting a deceptive instruction that tries to bypass the AI's safety features
  - **Human Error**
  - **Supply Chain Risks**
    - Third-party plugins and APIs.

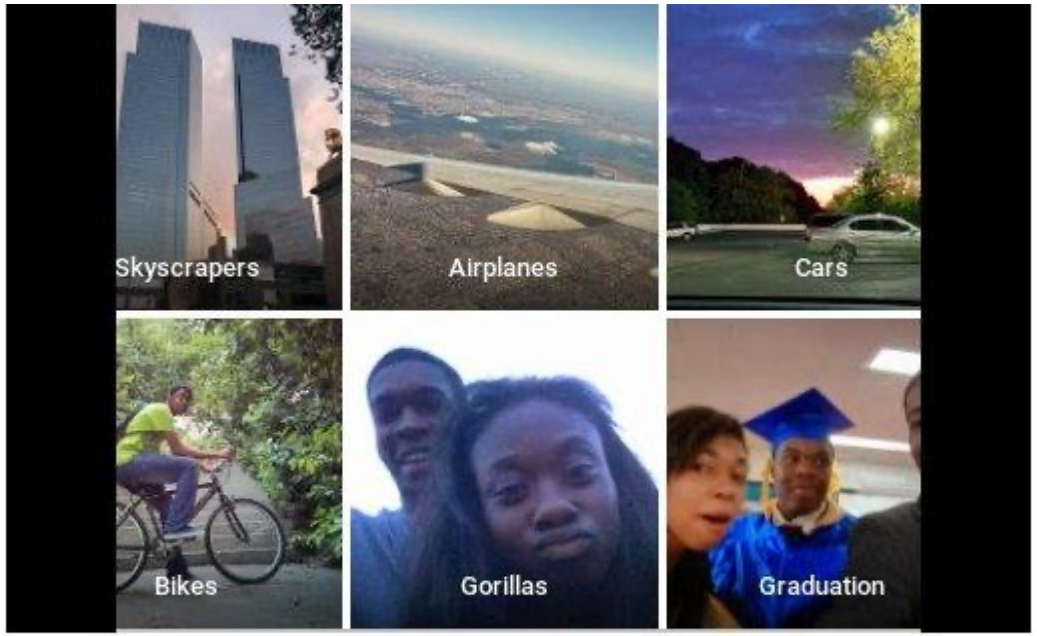


Input Risks: **Data Privacy & Security**

- **Discussion**
  - Besides the emergency scenario, what other types of data in your daily work at NISR would be absolutely forbidden to upload to a public AI platform? Why?
  - What types of data may be okay to upload? Why?

Input Risks: **Data Bias**

- AI systems are mirrors that reflect the biases present in the data they are trained on.



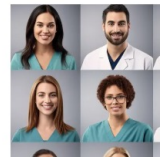
News | Article | August 7, 2024

AI has trouble picturing physicians that aren't White men

Author(s): Richard Payerchin

'Striking' lack of diversity when researchers use artificial intelligence to generate images of doctors.

Artificial intelligence (AI) thinks American physicians look mostly like White men, even as medicine diversifies with doctors who are women and are Asian, Black and Latino, according to a new study.



Can the criminal justice system's artificial intelligence ever be truly fair?

Computer programs used in 46 states incorrectly label Black defendants as "high-risk" at twice the rate as white defendants

Processing Risks: Inside the AI's "Mind"

Hallucinations & Factual Inaccuracy

- Generative AI models are designed to generate plausible-sounding text; they are not databases of facts.

Your Choices

You can control and access some of your Personal Data directly through settings. For example, you can manage your chat history. Should you choose to do so, you may also delete your chat history via your settings.

If you choose to delete your account, you will not be able to reactivate your account or retrieve any of the content or Personal Data in connection with your account.

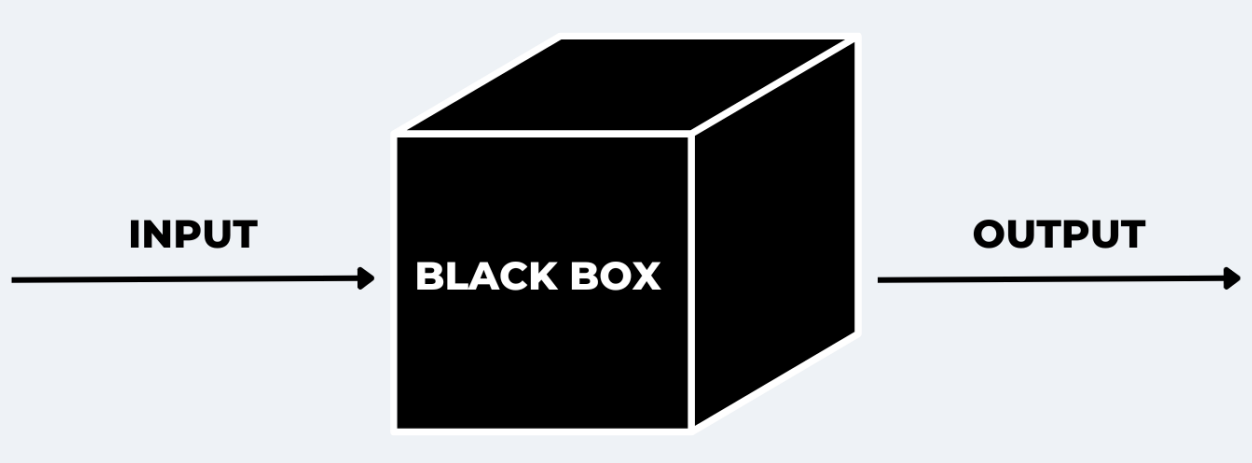
A note about accuracy: Services like DeepSeek generate responses by reading a user's request and, in response, predicting the words most likely to appear next. In some cases, the words most likely to appear next may not be the most factually accurate.

For this reason, you should not rely on the factual accuracy of Output from our models. If you notice that Output contains factually inaccurate Personal Data about you and you would like to request a correction or removal of the Personal Data, you can submit these requests through privacy@deepseek.com and we will consider your request based on applicable law and the technical capabilities of our models.

Processing Risks: Inside the AI's "Mind"

The "Black Box" & Lack of Explainability

- We can see the input and the output, but we cannot easily understand the internal reasoning process that led from one to the other.
- Any of the three components of a machine-learning system can be hidden, or in a black box.
  - a set of algorithms
  - training data
  - a model



Application Risks: The Societal Impact

Algorithmic Control & Power Dynamics

- Algorithms can be used to manage, monitor, and control people in ways that are subtle and pervasive.

- Case Study: The Food Delivery Riders
  - The AI system which optimized delivery times is constantly learning. It learns that riders can navigate traffic in certain ways, so it shortens the allotted delivery time.
  - Over 3 years, the average delivery time in China was reduced by 10 minutes
  - In Shanghai in 2017, one delivery rider was killed or injured every 2.5 days.
- The algorithm's objective function is a form of power.

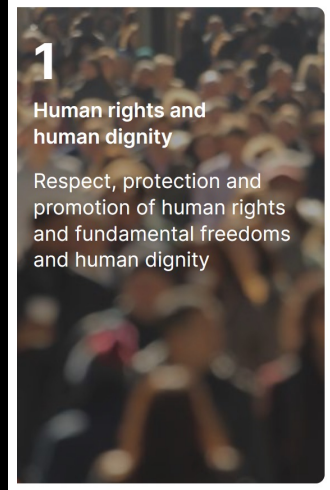
Application Risks: The Societal Impact

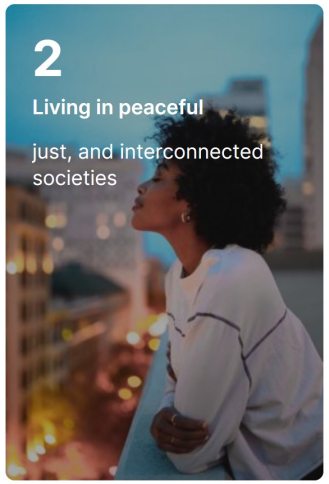
Over-reliance

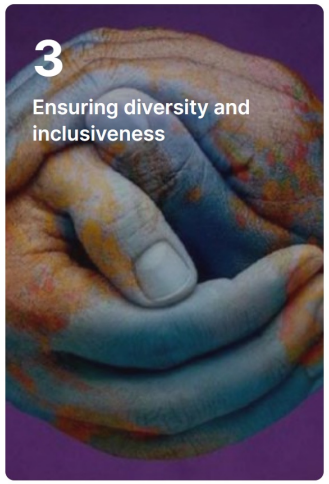
- We must ensure that AI is a tool that augments our intelligence, not a crutch that replaces it.

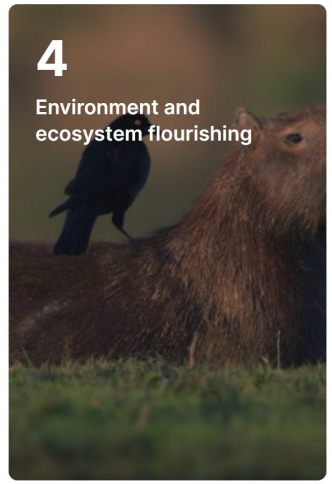
Mitigating the Safety and Ethical Risks of AI

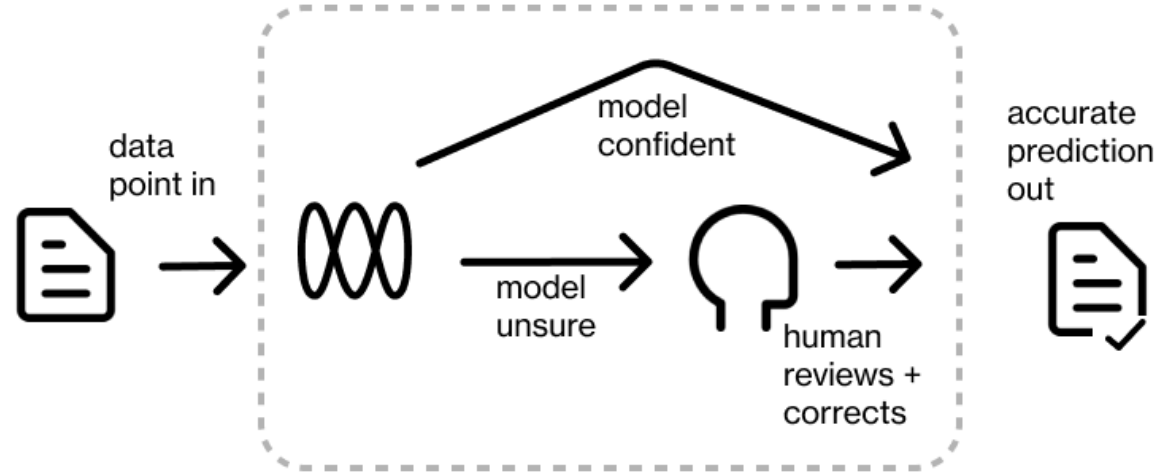
- Foundational principles based on a global consensus
- Strategies for upholding data privacy and security
- Strategies for mitigating ethical risks











Workers-in-the-Loop AI deployment

Humanloop

Foundational Principles based on a Global Consensus

- The OECD AI Principles

Values-based principles

The OECD AI Principles promote use of AI that is innovative and trustworthy and that respects human rights and democratic values. Adopted in May 2019, they set standards for AI that are practical and flexible enough to stand the test of time.

Inclusive growth, sustainable development and well-being +

Human rights and democratic values, including fairness and privacy +

Transparency and explainability +

Robustness, security and safety +

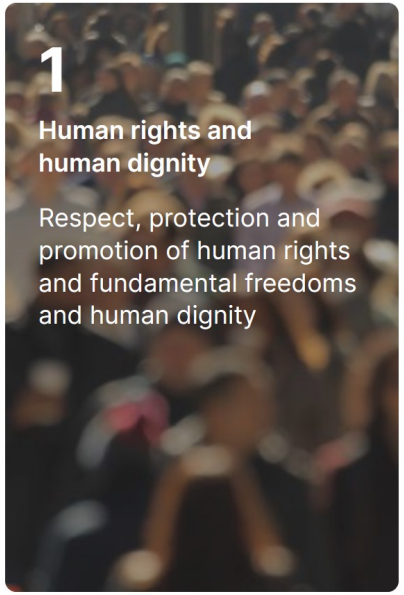
Accountability +

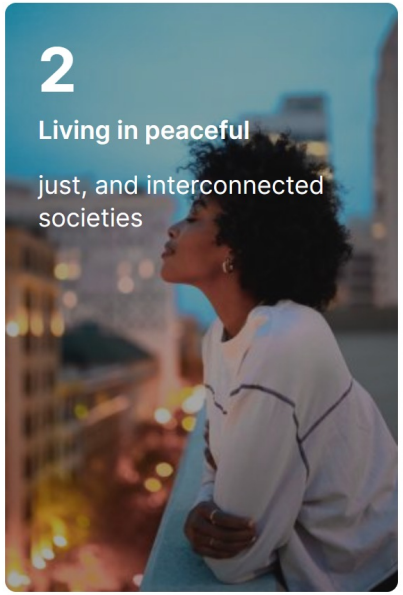
Foundational Principles based on a Global Consensus

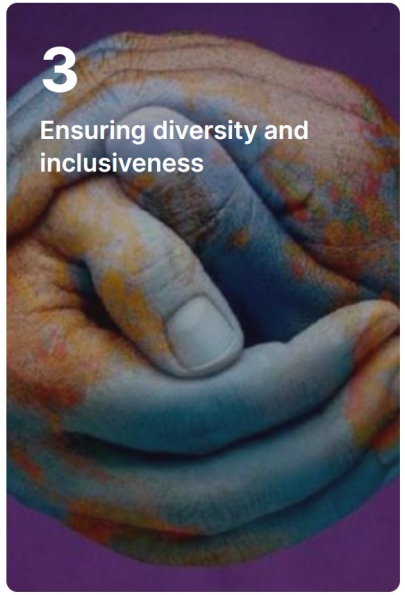
- The UNESCO Recommendation on the Ethics of AI

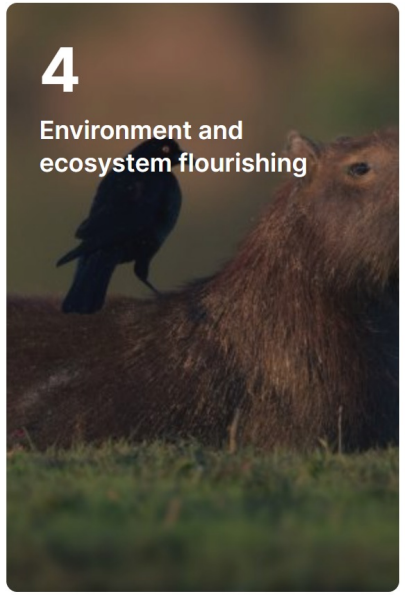
Four core values

Central to the Recommendation are four core values which lay the foundations for AI systems that work for the good of humanity, individuals, societies and the environment:









Foundational Principles based on a Global Consensus

- What would you say, be the foundational principles of AI usage?
- **Comparative Analysis of International AI Principles**

Principle/Theme	**OECD AI Principles**	**UNESCO Recommendation on the Ethics of AI**
Accountability	Is in the listed principles.	Requires auditable and traceable systems. There should be oversight, impact assessment, and due diligence mechanisms to ensure responsibility.
Transparency & Explainability	Is in the listed principles.	The ethical deployment of AI depends on transparency and explainability (T&E). The level of T&E should be appropriate to the context.
Fairness & Non-Discrimination	"Human rights and democratic values, including fairness and privacy"	AI actors should promote social justice, fairness, and non-discrimination, taking an inclusive approach to ensure benefits are accessible to all.
Data Governance & Privacy	"Human rights and democratic values, including fairness and privacy"	Privacy must be protected and promoted throughout the AI lifecycle. Adequate data protection frameworks should be established.
Human Oversight	Implement mechanisms and safeguards, such as capacity for human agency and oversight, to address risks.	Member States should ensure that AI systems do not displace ultimate human responsibility and accountability. Life and death decisions should not be ceded to AI.
Robustness & Security	Is in the listed principles.	Unwanted harms (safety risks) and vulnerabilities to attack (security risks) should be avoided, prevented, and eliminated throughout the lifecycle.

Strategies for Upholding Data Privacy and Data Security

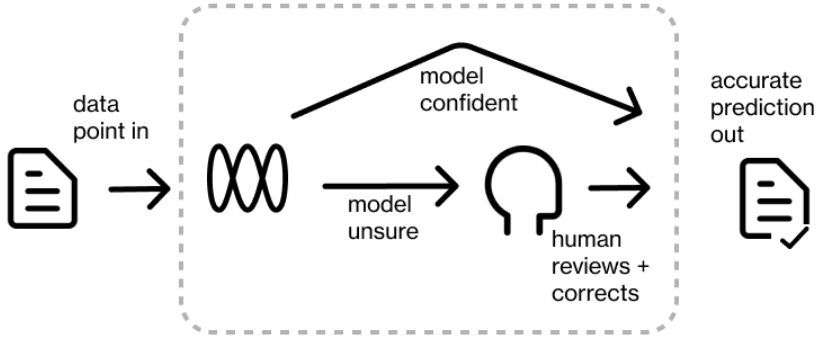
Keep Data In-House

1. Prioritize local deployment
2. Anonymize and De-identify Data Before Use
   - Replacing real, detailed information with fuzzy information
   - Masking
   - Tokenization
     ① The Original Data (“The Cash”)
     ② Generate a Token (“Exchanging Cash for a Chip”)
     ③ Use the Token (“Using the Chip at the Tables”)

Strategies for Mitigating Ethical Risks

Human-in-the-loop

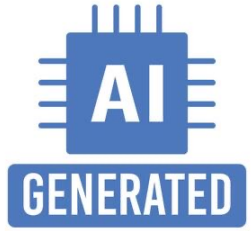
1. Pre-Screen Data for Potential Bias
2. Verify, Then Trust
   - ➢ Ask for source, check and cite the primary source
3. Implement Diverse Review



Workers-in-the-Loop
AI deployment

Humanloop

Develop Clear Governance & Accountability



1. Transparency & Labelling
   - *China: "Provisions on the Management of Deep Synthesis" mandates that any service provider whose technology generates "deepfakes" must add a prominent label to the content*
   - *EU: "AI act", requires labelling AI generated contents*
2. Assigning Responsibility

Summary

Identify the Risks

1. Data Privacy & Security
2. Data Bias

3. Hallucinations & Factual Inaccuracy
4. The "Black Box" & Lack of Explainability

5. Algorithmic Control & Power Dynamics
6. Over-reliance

Mitigate the Risks

- Accountability
- Transparency & Explainability
- Fairness & Non-Discrimination
- Data Privacy
- Human Oversight
- Robustness & Security

1st: Keep the Potential Risks in Mind!

Keep Data In-House

1. Prioritize local deployment
2. Anonymize and De-identify Data Before Use

Human-in-the-loop

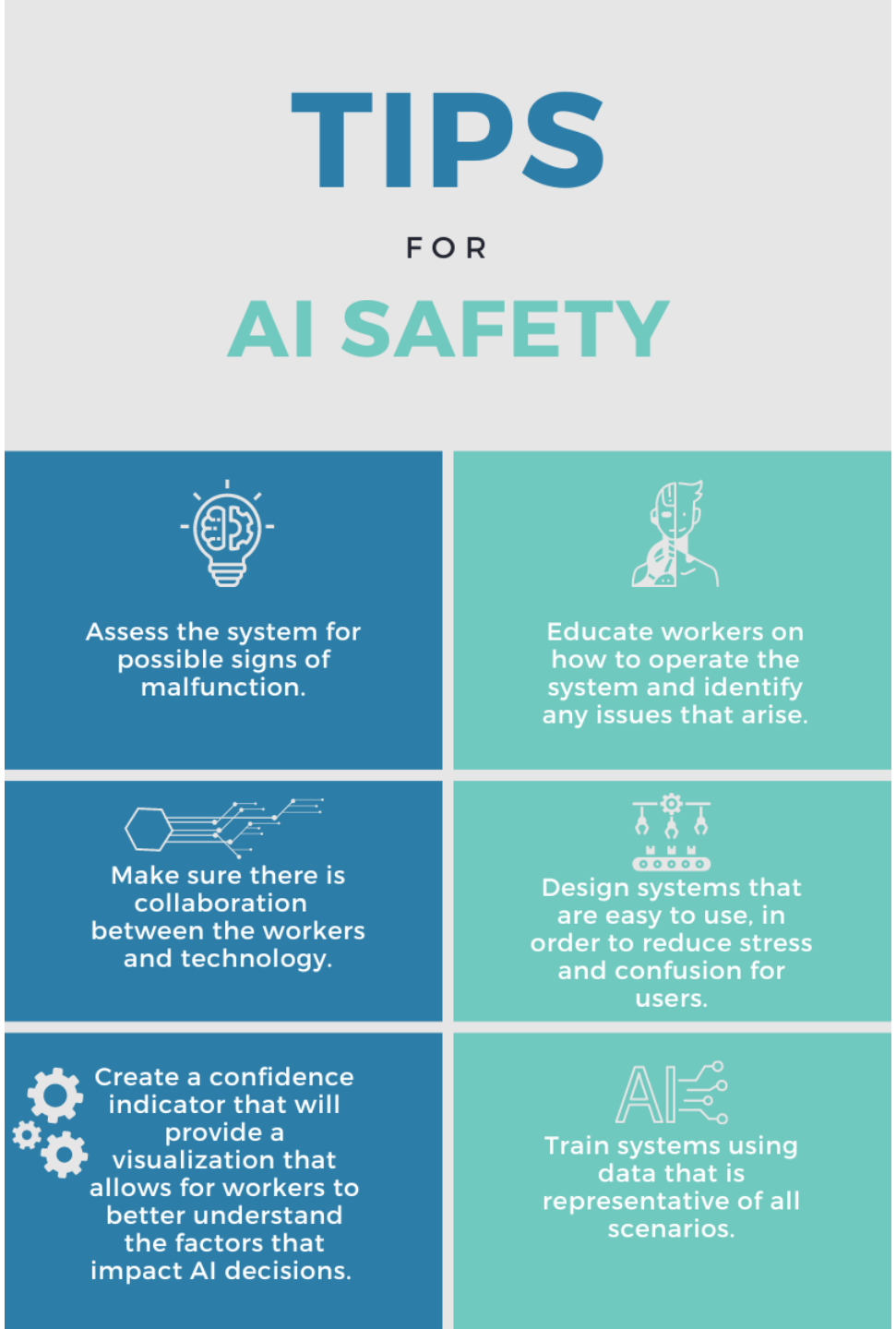
1. Pre-Screen Data for Potential Bias
2. Verify, Then Trust
3. Implement Diverse Review

Develop Clear Governance & Accountability

1. Transparency & Labelling
2. Assigning Responsibility

Challenge

- Develop a *Checklist for Using AI Responsibly and Effectively* for your own office



Thank you!

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY