Presentation for Japanese Ministry of Internal Affairs and Communications (Nov 25, 2020)

Putting our AI Principles into practice

Google

Google's AI Principles

AI should:

1. be socially beneficial
2. avoid creating or reinforcing unfair bias
3. be built and tested for safety
4. be accountable to people
5. incorporate privacy design principles
6. uphold high standards of scientific excellence
7. be made available for uses that accord with these principles

Applications we will not pursue:

1. likely to cause overall harm
2. weapons or those that direct injury
3. surveillance violating internationally accepted norms
4. purpose contravenes international law and human rights

Principles that remain on paper are meaningless.

-Sundar Pichai, *Financial Times*, Jan. 2020

Our goal: Earn and maintain our user's trust

Knowledge Base

Develop Research, best practices, resources

Thoughtful Products

Proactive, end-to-end improvements

Dialogue

Share our learnings and continue to seek perspective & feedback from diverse experts

How we put our Principles into practice

Culture and Education:

Training, resources and workshops

Tools, Techniques & Infrastructure:

Data, models, testing, publications

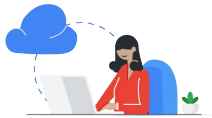
External Engagement:

Conferences, consultations

Structures and Processes:

Sensitive topics guidance, reviews, escalation

Culture and Education



Tech Ethics Trainings

Link

Human-centered Design workshops

ML Fairness Trainings

Link

Issue Spotting Training

People + AI Guidebook

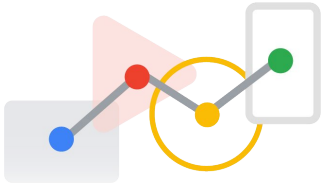
Online Guides and Resources

Link

Machine Learning for Policy Leaders

Tools, Techniques and Infrastructure

Data



Facets:
open source tool to analyse datasets

Crowdsource:
more diverse data

Data cards:
“nutrition labels” for datasets

ML Models

TensorFlow Lattice:
open source library to add in policy constraints

GDIQ:
Building models that help to detect bias

Model cards:
“nutrition labels” for models

Assessments

Fairness Gym, Fairness Indicators, What-If Tool, etc.:
assessments of different fairness goals

Adversarial Testing:
fairness testing and monitoring

External Engagement

BSR®

THE CENTRE FOR internet & society

Future of Humanity Institute

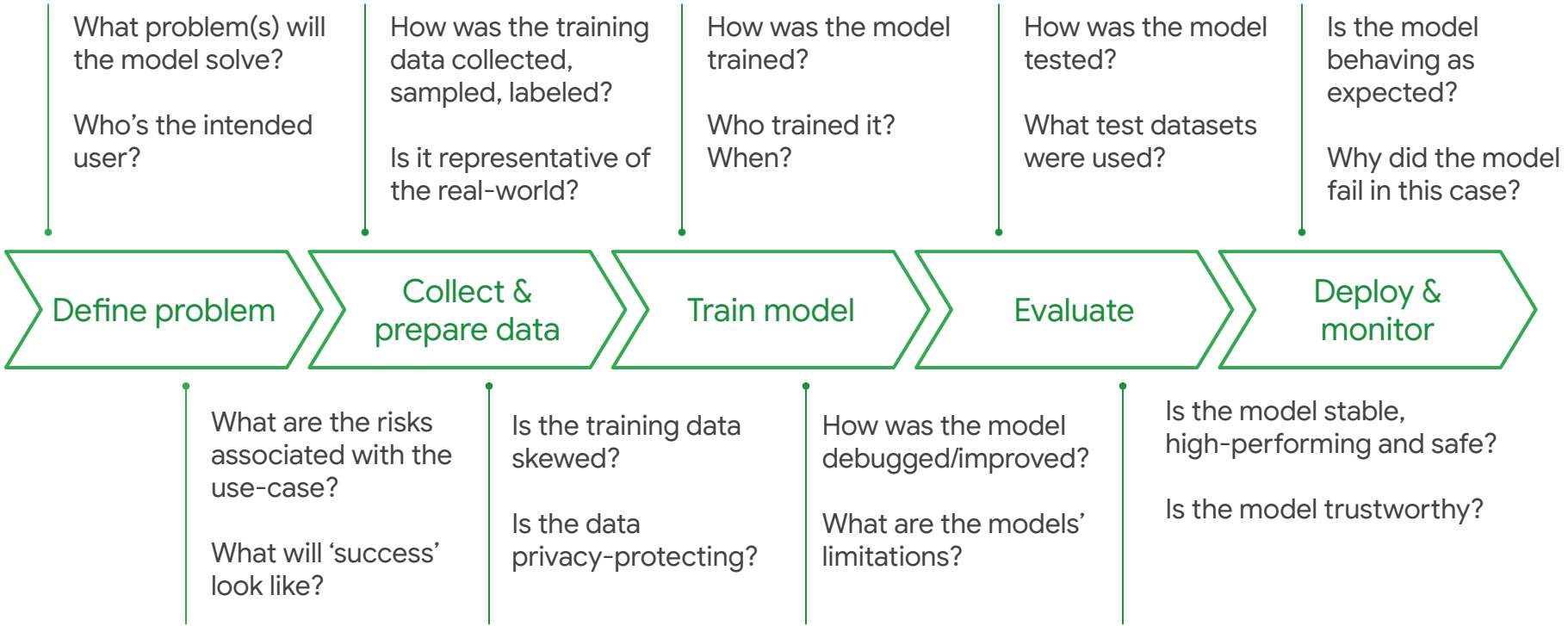
UNIVERSITY OF OXFORD

ISO

MLPerf

OECD.AI Policy Observatory

Building responsible AI requires answering hard questions across the ML lifecycle



AI Governance Structures and Processes

Across Google

Product Areas

Senior Executive Council & Advisors

Responsible Innovation Team

Consulted for:

- Sensitive cases and topics
- Tech that might affect multiple product areas

Serves as central hub for:

- Leading AI Principles evaluations, guidance
- Subject matter expertise in AI ethics, socio-technical research, human rights, law, content, etc.

Take ownership for:

- Implementing and managing dedicated processes unique to their needs, with guidance from central team

The AI Principles Review Process

Intake

Any Googler can request a review

Proactive pipeline for reviews

Central review team applies relevant AI Principles as ethical frameworks

Internal product, ethics, fairness, security, privacy, and other experts offer specific guidance

Analysis

Reviewers consider scale, scope of likely benefits and harms

Reviewers ask questions that reflect the AI Principles

Reviewers look for precedents to apply, similar to a case law process

Adjustment

Product/research team engages in specific technical evaluations

If necessary, reviewers consult with experts on mitigation strategy

Product/research team adjusts approach based on reviewers' mitigation guidance

Decision

IF challenging issues arise that can affect multiple products, a senior council of Google executives makes the decision to pursue or not pursue

OR central review team decides

Final decision can become a precedent; product/research team acts on mitigation strategy

Note: Each review is unique. This summary is intended as a high-level representation of the current process.

Text-to-speech: AI Principles Review outcomes

Approved:

Research paper

Publishable with cautionary language

Approved:

In products

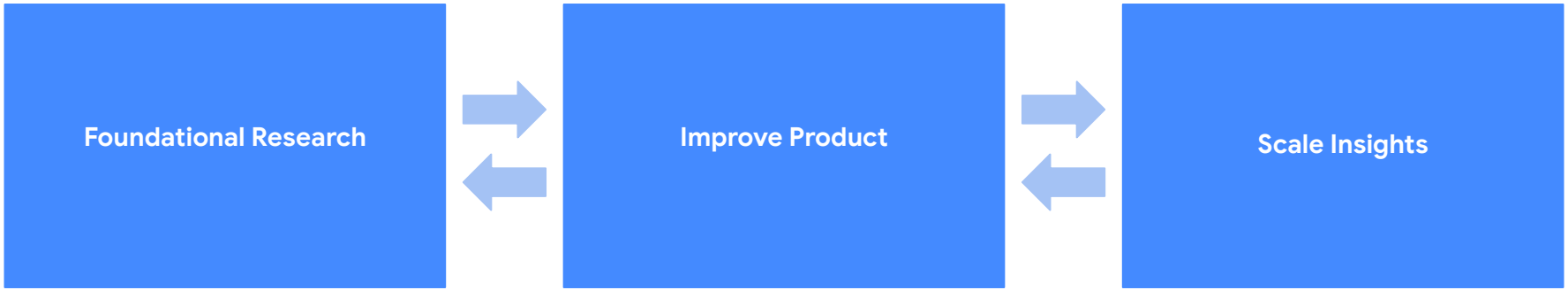
TTS acceptable, but only with user consent

Not approved:

Open source

Releasing the TTS model openly can risk malicious use

Good governance requires constant iteration



Questions?