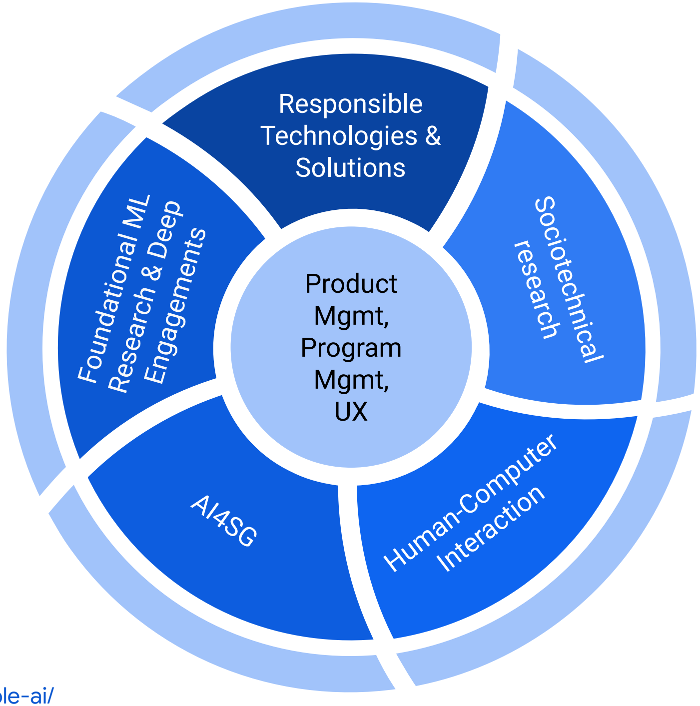
Responsible AI for Generative Models

Kathy Meier-Hellstern, PhD
Principal Engineer and Director, Google Research,
Responsible AI and Human-Centered Technology
kathyhellstern@google.com

February 13, 2023

Responsible AI and Human-Centered Technology

We conduct research and develop methodologies, technologies, and best practices to ensure AI systems are built responsibly.



https://research.google/research-areas/responsible-ai/

We are in the midst of a technology disruption

https://imagen.research.google/

There have been significant advances in generative models

Language: GPT-3, GLaM, Gopher, PaLM, Chinchilla, ChatGPT
Text-to-Image: DALL-E2, Stable Diffusion, Imagen, Parti

People are simultaneously excited about the new possibilities, and concerned about the possibilities for harm

The models have been broadly shared, with varying levels of safeguards

A small cactus wearing a straw hat and neon sunglasses in the Sahara desert.

Translation

Entertainment

Creativity

Automation

How do we build on those advances in a responsible manner?

Incites violence

Hallucination

Disinformation

NSFW

Leak
private data

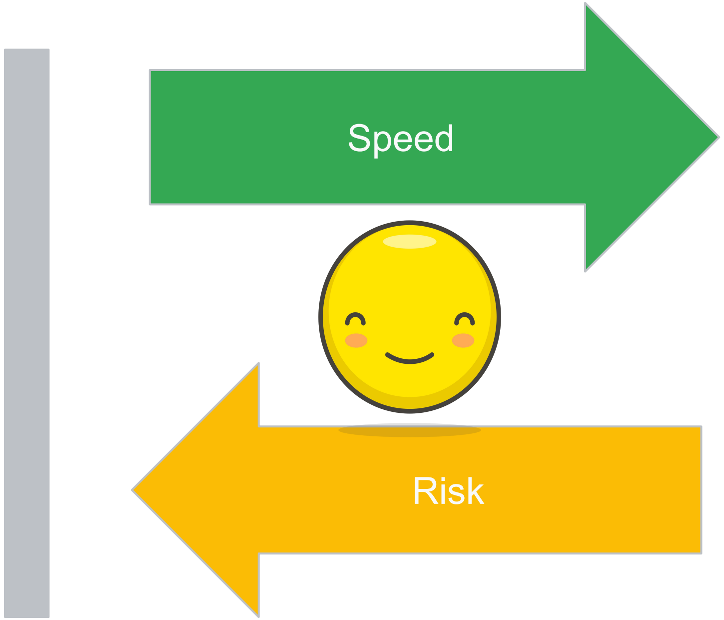
Deepfakes

Stereotypes
& Bias

What does Responsible AI look like in this environment?

Speed in a risk-appropriate manner while maintaining "hard lines"

- **Grounded** in a deep understanding of sociotechnical risks, harms and benefits
- **Responsible Data Practices**
- **Adaptive** for nuanced model generation and understanding
- **People-centric** for users with minimal AI expertise and programming
- **Feedback-enabled** to identify and mitigate concerns



Our Work is Grounded in our AI Principles

Google AI Principles

AI should:

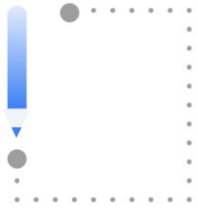
1. be socially beneficial
2. avoid creating or reinforcing unfair bias
3. be built and tested for safety
4. be accountable to people
5. incorporate privacy design principles
6. uphold high standards of scientific excellence
7. be made available for uses that accord with these principles

Applications we will not pursue:

1. likely to cause overall harm
2. technologies primarily intended to cause injury
3. surveillance violating internationally accepted norms
4. purpose contravenes international law and human rights

https://ai.google/principles/

The AI Principles Review Process

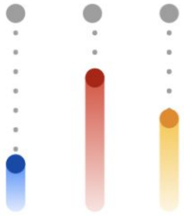


1. Intake

Gather information.
Apply ethical frameworks
and precedents.
Engage experts

2. Analysis

Consider scale, severity,
likelihood and ability to
mitigate benefits and
harms



3. Adjustment

Product/research team
adjusts approach based on
mitigation guidance.
Escalate if needed

4. Decision

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

Translated Wikipedia Biographies dataset: AI Principles Review

AI Principle 2:

Avoid creating or reinforcing unfair bias.

The Translate researchers built a new model that incorporates context from surrounding sentences or passages to improve gender accuracy when personal pronouns are translated.

AI Principle 4:
Be accountable to people.

AI Principles reviewers recommended that the researchers publish a data card, which is a structured document offering details about how the dataset was created and tested.

Translated Wikipedia Biographies

English -> Spanish (516 KB)
English-> German (517 KB)

The Translated Wikipedia Biographies dataset has been designed to evaluate gender accuracy in long text translations (multiple sentences or passages). The set has been designed to analyze common gender errors in machine translation like incorrect gender choices in anaphora resolutions, possessives and gender agreement.

PUBLISHER(S)

Google LLC

FUNDING

Google LLC

INDUSTRY TYPE

Corporate - Tech

FUNDING TYPE

Private Funding

DATASET AUTHORS

Anja Austermann, Google Michelle Linch, Google
Romina Stella, Google Kellie Webster, Google

DATASET CONTACT

translate-gender-challenge-sets@google.com

DATASET PURPOSE(S)

Testing

KEY APPLICATION(S)

Machine Translation, Gender Accuracy

ACCESS COST

Open Access

PRIMARY MOTIVATION(S)

Study gender accuracy in translations beyond the sentence in demographic and occupations diversity for fairness research.

INTENDED AND/OR SUITABLE USE CASE(S)

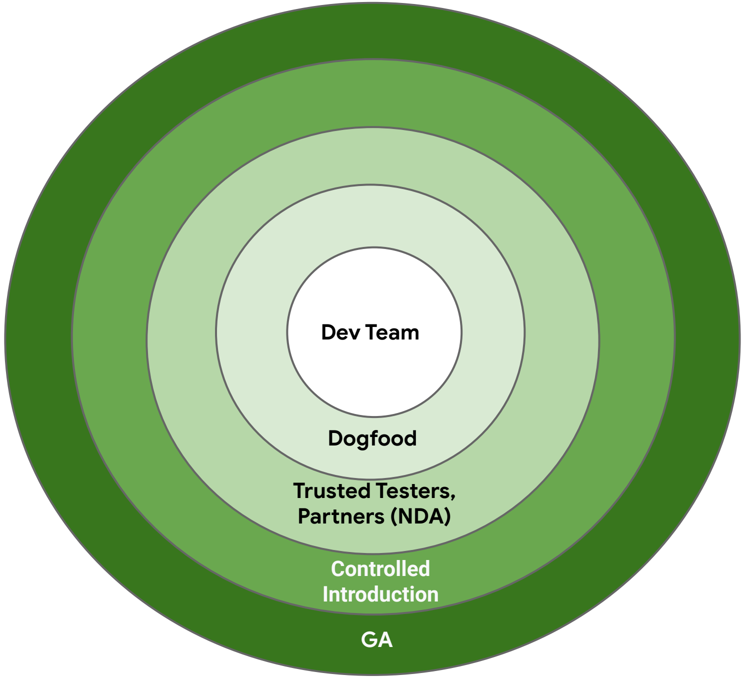
To evaluate gender accuracy on translations beyond the sentence (multiple sentences or passages). The set is focused on the presence of this specific linguistic

AI Principle 6:
Uphold high standards of scientific excellence

The Translate researchers decided to share the dataset publicly in order to support long-term improvements on ML systems focused on pronouns and gender in translation.

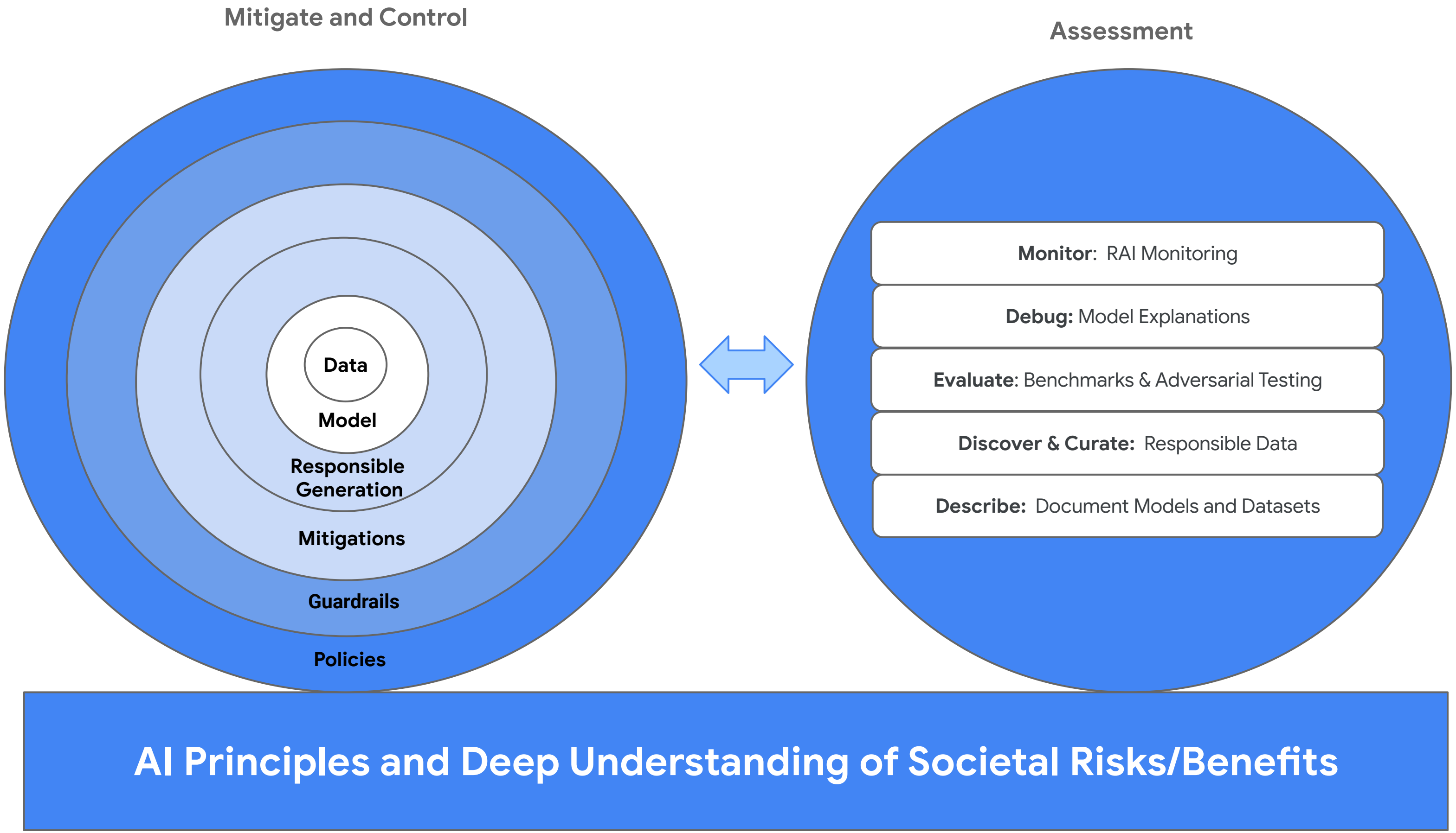
RAI Risk Management Vision
“Inside Out” and “Outside In”

Managed Risk Exposure - "Inside out"



- Gradual risk exposure, higher risks acceptable for "trusted users"
- Launch criteria matched to risk tolerance
- User Feedback to capture failures for future mitigation and testing
- Risks can be discovered and mitigated before they reach the general public

RAI Mitigation and Control - “Outside In”



Google

Example Research Areas

Example research focus areas

1. Sociotechnical research on risks, harms, benefits of Generative AI.
2. Responsible Data - High-quality RAI-focused datasets and agile classifiers for testing, fine tuning and reinforcement learning; new methods for working with human raters.
3. Controlled Generation to responsible steer model output
4. Scalable adversarial testing to uncover risks
5. New people-centric tools, services, and information that empower people to easily create, prototype, and control AI, with minimal AI expertise and programming

1. Sociotechnical risks, harms, and benefits to inform policies

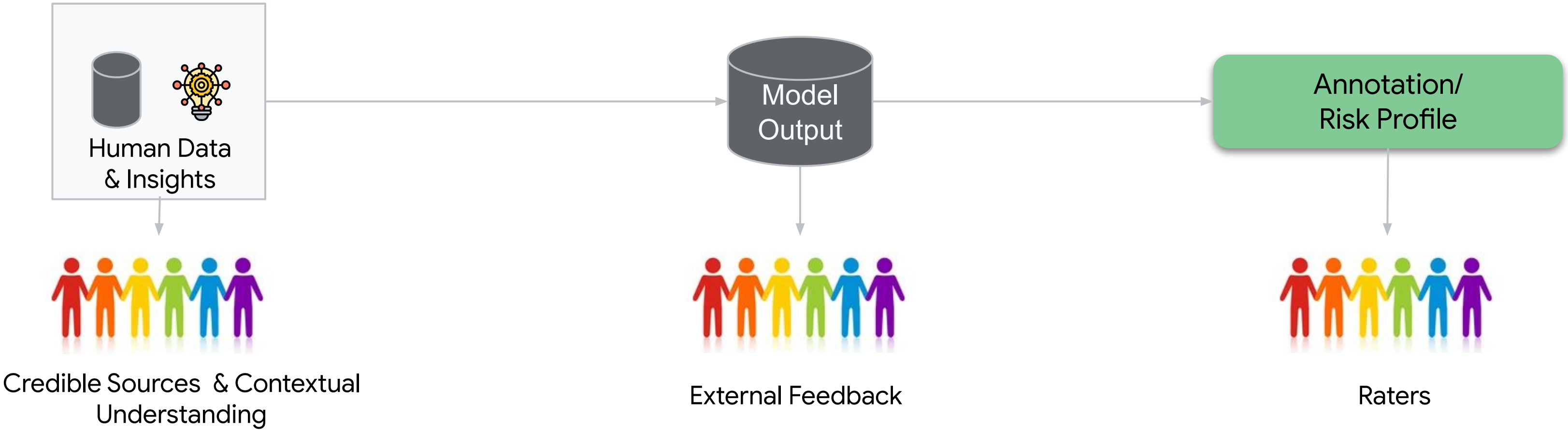
The foundation on which everything is built:

- **RAI risks** and **harms assessments** are grounded in:
  - **Evidence-based** and **community-centered research**
  - **Cultural context**
  - **Region specific foundational research**
- Hot topics:
  - Responsible **human–AI interactions** (account for **users' mental models** and **abilities**)

Community-based Research https://research.google/pubs/pub51151/
Toward Culturally-inclusive AI https://ai-cultures.github.io/
RAI User Experience https://research.google/pubs/pub52061/

Google

2. Responsible Data with Humans in the Loop



- Leverage cultural knowledge, external authoritative sources & datasets for critical domains
- Research for topical diversity per policy area, e.g., skin cancer for medical advice
- Public data challenges to collect adversarial examples (MLCommons DataPerf)

- Experts in education, environment, health, criminal justice
- Feedback and data from experts, e.g., practitioners in specific industries, in Africa
- Community-based/participatory work to bringing broader sociotechnical perspectives to guide evaluation and mitigation

- Quality of human annotated data, rater diversity
- Raters are not anonymous proxies for users
- Experts are not able to cover the spectrum of diverse opinions
- Experts & Raters don't agree among each other

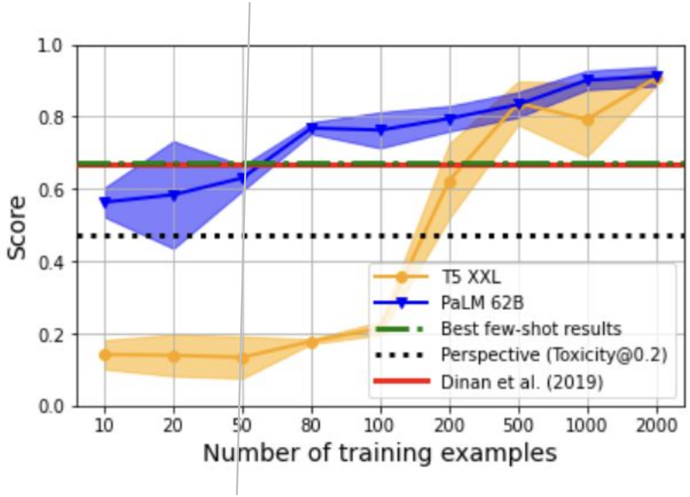
https://arxiv.org/abs/2207.10062

Google

2. Agile Classifiers - Parameter Efficient Tuning

Safety is a moving target

- We need quick ways to train safety classifiers
    - With small amounts of data (<<1000 examples) to be agile (can't get >>10k examples quickly)
- Prompt-tuned LLMs make for good classifiers with small amounts of data



From: Parlai single adversarial set (24,000 training examples)

Parameter Efficient Tuning: https://arxiv.org/abs/2205.05638
arxiv paper pending: "Toward Agile text classifiers for everyone"

Google

3. Controlled Generation for Improved Safety

Post-hoc filtering/reranking is not enough!
How do we generate better responses?

- We need mechanisms to guide generation towards safe and desirable outcomes
  - The guiding mechanism should introduce negligible latency while being effective.
- Training and inference time improvements can shift generation toward higher-quality, safer responses.
- Controlled generation is an effective mechanism to draw desirable outcomes in a streaming fashion.
  - e.g., prompt engineering, control tokens, prefix safety scores

Control Tokens: https://arxiv.org/pdf/2009.06367.pdf
Prompt-based prototyping: https://dl.acm.org/doi/abs/10.1145/3491101.3503564
Chain-of-Thought Prompting: https://arxiv.org/abs/2201.11903

4. Adversarial Testing

Adversarial queries ...

... are likely to cause a model to fail in an unsafe manner (i.e. safety policy violations)

Adversarial queries ...

... cause errors that are easy for humans to identify, but difficult for machines to recognize.

Different flavors of “adversariality”

Explicitly Adversarial queries contain ...

Policy-violating language or express policy-violating points of view. E.g., slang.

Probing / attacks to trick or break the model into saying something unsafe, harmful or offensive.

Implicitly Adversarial

Innocuous queries that contain sensitive topics that are contentious, culturally sensitive, or potentially harmful.

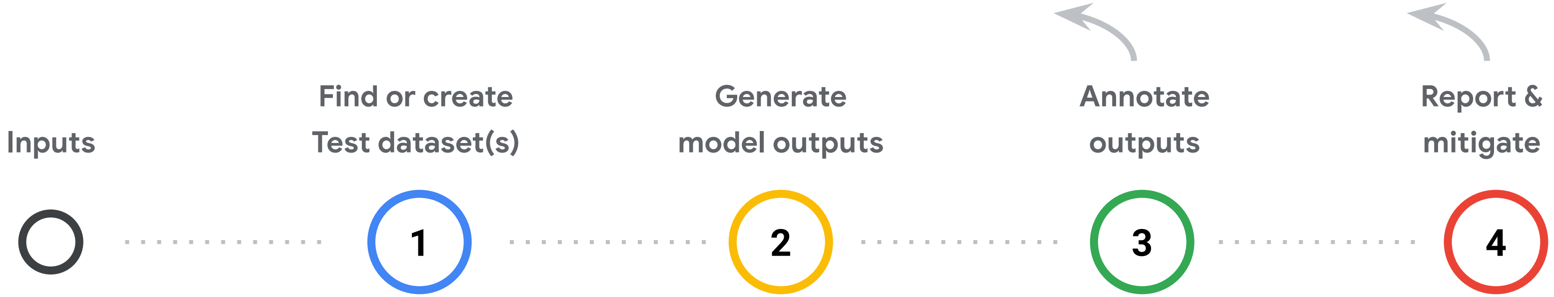
E.g., demographics, health, finance, religious holidays.

Why is adversarial testing useful?

1. Helps teams improve models & products by exposing current failure patterns to guide mitigation pathways.
   e.g., Fine-tuning, model safeguards / filters.

2. Informs product launch decisions by measuring risks that may be unmitigated.
   e.g., Likelihood the model will output policy-violating content.

Adversarial Testing Workflow



Adversarial Testing Workflow

Inputs

1. Find or create Test dataset(s)
2. Generate model outputs
3. Annotate outputs
4. Report & mitigate

Product Policy describing potential safety failure modes

Use-Cases: e.g., “write a blog post”, “summarization”

Diversity Reqts: Lexical, semantic, representation, etc

Adversarial Testing Workflow

Inputs

1 Find or create Test dataset(s)

2 Generate model outputs

3 Annotate outputs

4 Report & mitigate

1a. **Find** existing dataset(s)

1b. **Collect** queries (human-generated prompts)

1c. **Expand** dataset using data synthesis methods

1d. **Analyze** data quality & diversity

Adversarial Testing Workflow

Inputs

Find or create Test dataset(s) — 1

Generate model outputs — 2

Annotate outputs — 3

Report & mitigate — 4

2. Generate model output for all queries

Adversarial Testing Workflow

Inputs

1 Find or create Test dataset(s)

2 Generate model outputs

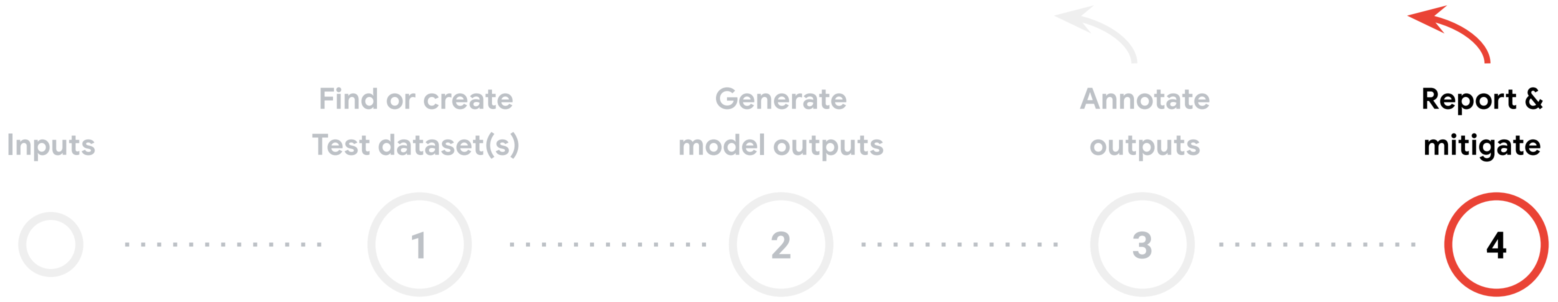
3 **Annotate outputs**

4 Report & mitigate

3a. Automatically annotate model outputs using safety classifiers

3b. Manually annotate model outputs using human raters

Adversarial Testing Workflow



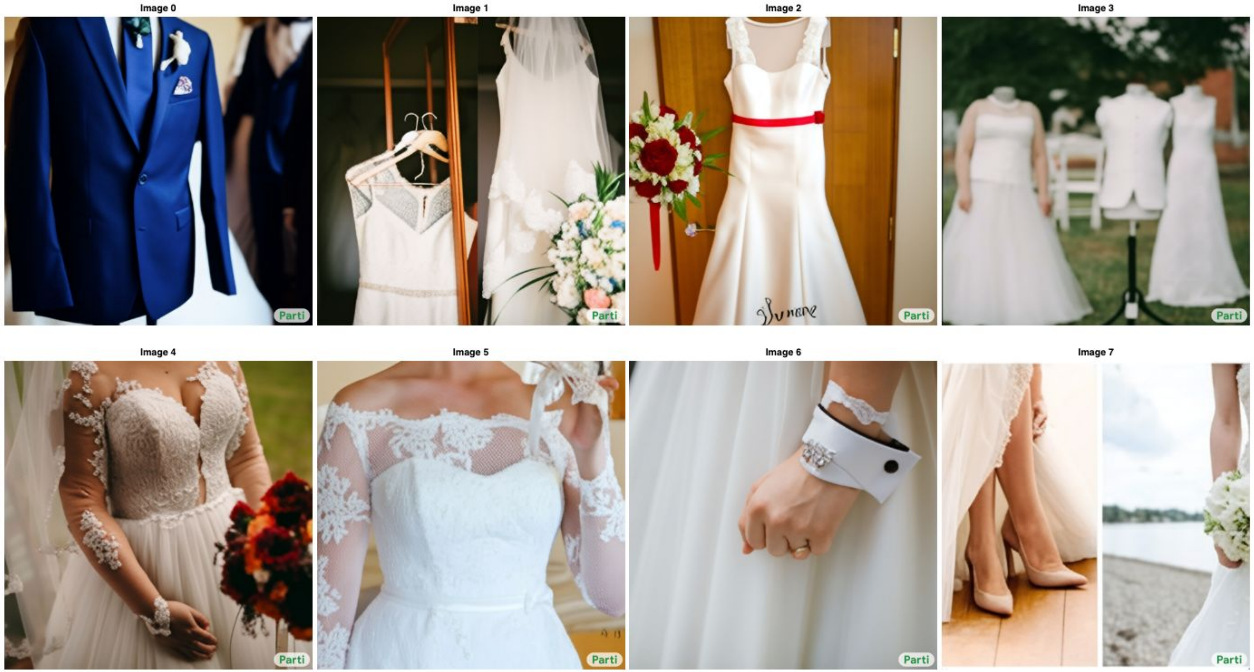
4a. Compute metrics & report results to decision-makers

4b. Guide model improvements based on test results

4c. Inform model safeguards (e.g. filters, blocklists) based on test results

Adversarial Testing Example - Safety and Fairness

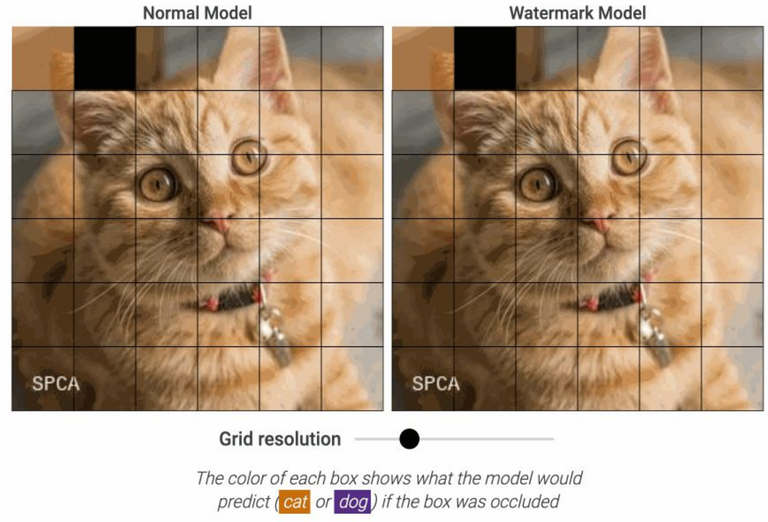
photo of clothing for a wedding



What Might Have Been - examples from Image Search

5. People-centric tools, services, and information

- **Explorables:** Interactive explorable visualizations that introduce key ideas and guidance to the research community,
- **Model Cards** and **Data Cards**: organize and communicate the essential facts of models and training data in a structured way
- **Learning Interpretability Tool:** an open-source platform for visualization and understanding of ML models
- **Know Your Data**: allows interactive qualitative exploration of models and big datasets
- **HCI Research for LLMs**: e.g. prompt-based prototyping for quick testing and learning



PAIR Guidebook: https://pair.withgoogle.com/guidebook/
Model Cards: https://modelcards.withgoogle.com/model-reports
Data Cards: https://dl.acm.org/doi/fullHtml/10.1145/3531146.3533231
Data Cards Playbook: https://sites.research.google/datacardsplaybook/
Know Your Data: https://ai.googleblog.com/2021/08/a-dataset-exploration-case-study-with.html
Prompt-based prototyping: https://dl.acm.org/doi/abs/10.1145/3491101.3503564

Putting it all together: What success looks like

- We have created tools and solutions that are innovative and flexible in striking the balance between moving with speed and being responsible in a risk-appropriate manner, and that can be used by development teams to quickly identify and remediate problems.
- We implement techniques for responsibly aligning model output with the principles and policies that have been established.
- Knowing that non-experts will be using our models, we create new ways to support them with easy access and tools and techniques for building responsibly.
- We have deeply explored the implications of these models on society, and can articulate the "hard lines" of responsibility, situations where we cannot compromise without violating our AI Principles.
- All products use the responsible AI tools by default as they become part of the technical infrastructure so that responsibility is baked into their applications.

Where to go for more information

Google AI Blog: https://ai.googleblog.com/

Google's AI Principles: https://ai.google/principles/

RAI-HCT Website: https://research.google/research-areas/responsible-ai/

Google Research, 2022 & beyond: Responsible AI Blog:
https://ai.googleblog.com/2023/01/google-research-2022-beyond-responsible.html

Thank You!