Google

2022 AI Principles Progress Update

This page is intentionally left blank.

Table of Contents

Introduction

This year, we've seen AI help people find useful information in the conversational way they speak to friends and family, thanks to responsible research breakthroughs in generative language models;[1] enable online translation across dozens of new languages;[2] foster the equity of genomic tests[3] and maternal-fetal health[4] across populations; and speed up damage assessments[5] after natural disasters for humanitarian relief and environmental efforts.

That's just a sampling of ways in which AI will have — is already having — long-term positive impacts on society across the business, healthcare, education, and sustainability sectors.

For much of the world, the flurry of research and development is delivering that future already. Where AI-enabled products and services are only now emerging, there is an opportunity — and a responsibility — to proactively identify and mitigate risks. Now is the time to establish effective frameworks that embody the technology, practices, values, and governance of responsible AI, because AI's wide-ranging impact naturally raises new questions about AI's governance, safety, fairness, and effect on equitable economic opportunities.

Because AI is core to Google products, we at Google ask these questions daily. We remain committed to sharing our lessons learned and emerging responsible innovation practices, drawing upon more than 20 years of using machine learning and more than a decade of AI research. Rooted in our near 25-year-old mission to organize the world's information and make it universally accessible and useful, Google's innovation strategy is to iterate on the process of innovation itself. This means that we create projects that not only exemplify engineering excellence, but from their earliest moments embody the human-centered values manifested in Google's AI Principles.

We do so by incorporating responsible practices for fairness, safety, privacy, and transparency early in developers' machine learning workflow and throughout the product development lifecycle. This principled approach to AI research and development is also practical: it can help avoid burning engineering cycles spent retrofitting technology if an issue emerges after launch or even much later. This aligns with our product excellence mantra to put the user first and with our focus on building for everyone.

Since we launched our AI Principles in 2018, we've built and tested an industry-leading governance process to align AI projects across the company with those Principles. We center our governance on three pillars:

1. **AI Principles**, which serve as our ethical charter and inform our policies
2. **Education and resources**, such as ethics training and technical tools to test, evaluate, and monitor the application of the Principles to all of Google's products and services
3. **Structures and processes**, which include risk assessment frameworks, ethics reviews, and executive accountability

This year, we expanded the central operations team for AI Principles implementation across Google's product development lifecycle, Responsible Innovation, and recently moved it into Google's company-wide Office of Compliance and Integrity for more centralized governance across all Google product areas. This is a milestone moment that reflects the growing maturity of our governance strategy.

We are working to complement these internal frameworks by working with a number of governments and organizations exploring concepts in AI governance. For instance:

- Organizations like the International Organization for Standardization (ISO)[6] and National Institute of Science and Technology (NIST)[7] are publishing AI management frameworks early next year.
- Singapore released its Model AI governance framework[8] and continues to talk with stakeholders in the financial services Veritas initiative.[9]
- India's Ministry of Electronics and Information Technology (MeitY) is considering Niti Aayog's proposed Responsible #AIForAll[10] to be incorporated into India's AI mission.
- Canada,[11] Brazil,[12] and South Korea[13] have all debated AI legislation, while the U.S.[14] the U.K.,[15] and Israeli[16] governments are working on their own guidance.

As our CEO has said, AI is too important not to regulate, and too important not to regulate well.[17] AI legislation and related principles and standards should help lower risks to people without unduly stifling innovation or undermining AI's promise for social benefit at the global level. And of course AI frameworks overlap with other important regulatory issues, including content safety, child safety, privacy, and consumer protection. A holistic approach will help keep new rules from impeding innovation and competition in AI and related emerging technologies.

We hope that sharing our progress and lessons learned on issues such as responsible AI, algorithmic transparency, privacy-enhancing technologies, and AI R&D supports the important progress being made across the global AI community

Google AI Principles

We will assess AI in view of the following objectives. We believe AI should:

1. **Be socially beneficial:** With the likely benefit to people and society substantially exceeding the foreseeable risks and downsides.
2. **Avoid creating or reinforcing unfair bias:** Avoiding unjust impacts on people, particularly those related to sensitive characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, ability and political or religious belief.
3. **Be built and tested for safety:** Designed to be appropriately cautious and in accordance with best practices in AI safety research, including testing in constrained environments and monitoring as appropriate.
4. **Be accountable to people:** Providing appropriate opportunities for feedback, relevant explanations and appeal, and subject to appropriate human direction and control.
5. **Incorporate privacy design principles:** Encouraging architectures with privacy safeguards, and providing appropriate transparency and control over the use of data.
6. **Uphold high standards of scientific excellence:** Technology innovation is rooted in the scientific method and a commitment to open inquiry, intellectual rigor, integrity and collaboration.
7. **Be made available for uses that accord with these principles:** We will work to limit potentially harmful or abusive applications.

In addition to the above objectives, we will not design or deploy AI in the following application areas:

1. Technologies that cause or are likely to cause overall harm. Where there is a material risk of harm, we will proceed only where we believe that the benefits substantially outweigh the risks, and will incorporate appropriate safety constraints.
2. Weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people.
3. Technologies that gather or use information for surveillance violating internationally accepted norms.
4. Technologies whose purpose contravenes widely accepted principles of international law and human rights.

Internal Governance and Operations

AI offers a unique range of risks along with its unprecedented benefits to the world. Chief among these are issues of technical safety, when AI systems do not function as engineers and designers have planned, and societal concerns, such as AI systems reflecting historical unfair bias.

We believe that through rigorous, structured operations, risks can be consistently identified and addressed, while acknowledging that even despite best efforts not all issues and harms can be identified in advance. Accomplishing these aims rests on three pillars: our AI Principles, Education and Resources, and Structures and Processes.

First Pillar: Our AI Principles

Google's AI Principles represent our first and foundational pillar. The Principles serve as our ethical charter and are a key component of Google's product excellence efforts. Consistent policies and responsible practices enable the structured governance that let us scale the practice of principled AI innovation.

Second Pillar: Education and Resources

The second pillar of AI governance at Google is our commitment to providing necessary education and resources. Our structured employee education programs — including in the onboarding for all new technical hires — offer a wide variety of resources that help Googlers learn the critical-thinking skills necessary to apply the AI Principles as frameworks for decision-making as early in the research and development process as possible. These programs are reaching an ever-growing portion of the Google population:

- Course completions of our self-study tech ethics training, offered to all employees across roles, product areas, and geographies, have increased 23% quarter over quarter for 2022.
- We have exceeded our 2022 goal of 20,000 full-time employees' engagement in our **Responsible Innovation Challenge**,[18] a set of interactive puzzles that test employees' recall of the AI Principles, which we launched last year.
- More than 34 groups have gone through our Moral Imagination workshops,[19] which help employees explore potential outcomes outside of their lived experience.

We continue to create innovative internal training programs to teach the skills needed to address emerging challenges. For example, this year, we launched a new, instructor-led course, **"Helping Users Navigate an AI World: Improving our products with an Explainability toolkit,"** geared toward user experience designers, product managers, and ML developers. It introduces the concept of Explainability, and showcases best practices in explaining how AI-powered products work. Internal ratings for the usefulness of the course's content average 4.5 out of 5 stars.

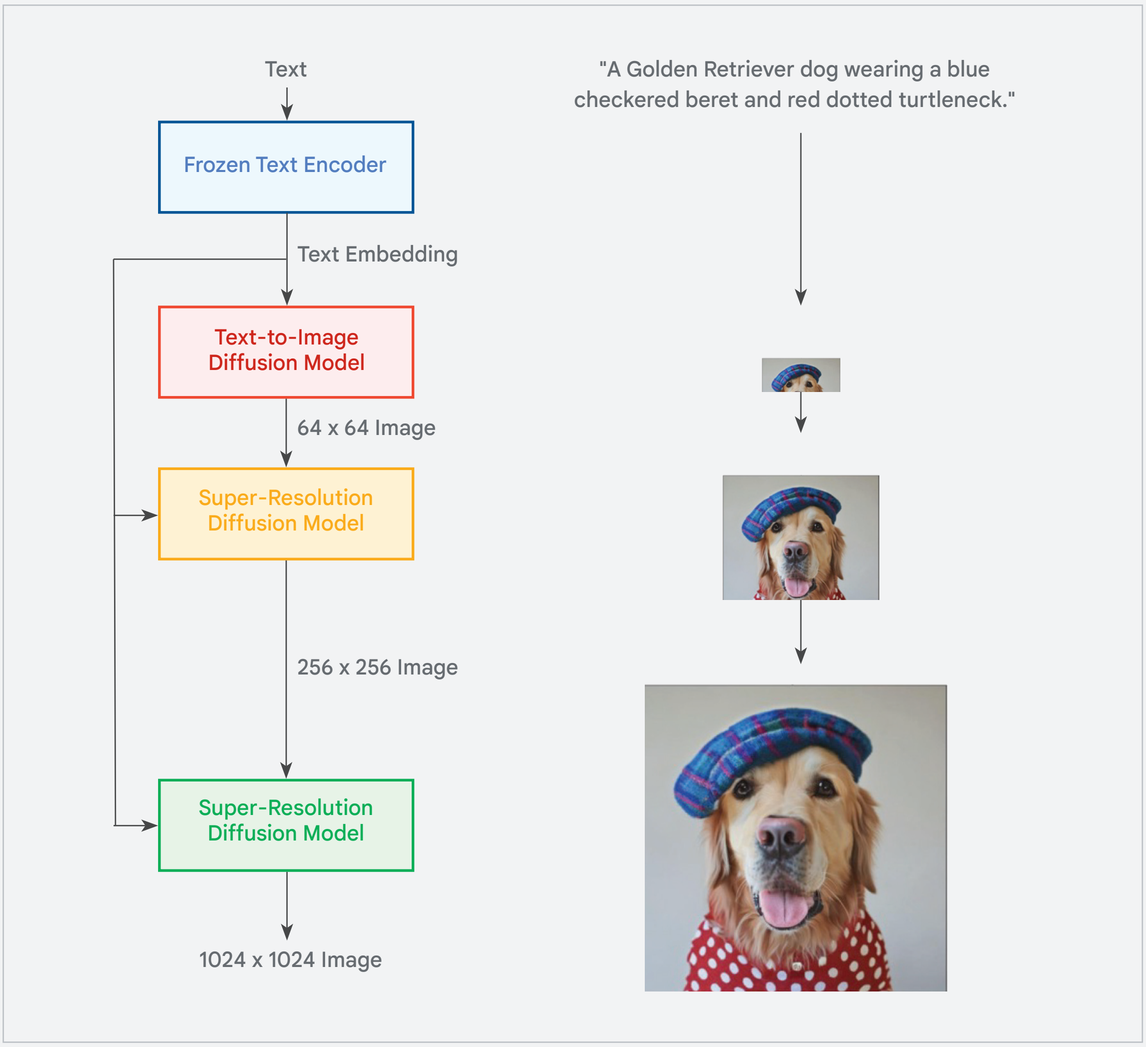
Our **cross-company AI Principles programs** are designed not only to educate but also to motivate and mobilize participants to take action. The programs we reported on last year grew in scope and reach in 2022. These programs are "20%" volunteer projects — Google's long-time practice of offering employees 20% of their working hours, or the equivalent of one work day per week, to dedicate to a passion project that could have business benefits for the company and value for our users, with the approval of their manager.

For example, the 20% **Principles Pioneers** program, which has an emphasis on people who represent points of view currently underrepresented in the technology industry, grew more than 460% this year — from 75 Pioneers in its pilot at the end of 2021 to now 423, at 98 Google international offices. Principles Pioneers constitute an internal community of trusted and trained employees who serve as AI Principles (AIP) advisors. They identify global fairness, harms, and human rights related concerns while stress testing AI-enabled products. They also supplement the work of the **ProFair** ("product fairness") team,[20] which also sits within our central AI Principles operations team. ProFair has the explicit specialization and goal of testing AI products for potential weaknesses and biases. By building centralized experience in red-teaming, we build upon the lessons of testing across different models.

This year we also launched a new internal program for senior managers and leaders: the **Executive AI Principles Ethics Fellowship**. This program includes educational workshops and training on the AI Principles for leaders across multiple product areas and geographies. It's based on our six-month **AI Principles Ethics Fellowship**, which we launched in 2020 and which has trained a diverse set of 50 employees from across 17 global offices and 15+ Employee Resource Groups to learn about responsible AI and contribute their perspectives to Google's AIP operations. During the fellowship, among other duties, fellows develop fictional, future scenarios for AI ethics challenges, addressing topics such as deep fakes and misinformation. Their hypothetical scenarios supplement a growing body of **responsible innovation case studies** that our AI ethics review teams draw upon as references when making decisions. The new Executive AI Principles Ethics Fellowship is tailored to business decision makers' needs. The inaugural cohort consisted of sixteen executives across ten product areas, including Cloud, Devices and Services (hardware), and YouTube.

AI Principles Case Studies

Case study: Text-to-image generation models



The Challenge

Parti[21] and Imagen[22] are Google's text-to-image generation models. These models allow users to provide a text prompt specifying an image, and that image is then generated directly from the text. AI Principles reviewers recognized that allowing the free generation of human images enables many significant potential harms (for example:

deep fakes, images of interpersonal violence, images conveying harmful stereotypes, pornographic images). Consequently, the decision was made to block the ability to generate people in initial testing, even internally. As the internal launch progressed, users expressed disappointment that they could not generate people with our models — especially because similar non-Google models allow for the generation of images of people. Researchers then tried to find a safe way to allow at least some images of people to be generated. To accomplish this, they employed scaled adversarial testing, which enabled the team to (1) generate 20,000 synthetic images of human faces, and (2) check to see whether they could use existing ML classifiers to selectively block the generation of photorealistic human faces — that is, faces that might belong to or resemble real people.

Researchers were able to show that an existing face classifier works well even on generated imagery. With this data in hand, they were able to loosen the prohibitive "no people allowed" rule, and began allowing the generation of imagery of people — so long as no photorealistic face is present in the image. However, while this mitigation helped to reduce the risk of deep fakes or confusion with a real person, there remained a risk of other harms such as unfair stereotyping.

The Approach

With guidance from the central Responsible Innovation team, Google is exploring a framework for responsible externalization that balances the value of external auditing with the risks of unrestricted open-access. The data requirements of text-to-image models have led researchers to rely heavily on large, mostly uncurated, web-scraped datasets. While this approach has enabled rapid algorithmic advances in recent years, datasets of this nature often reflect social stereotypes, oppressive or biased viewpoints, and derogatory or otherwise harmful associations to marginalized identity groups.

We've run dedicated rounds of adversarial testing to find flaws in the model. We enlisted expert red teaming members — product experts who intentionally stress test a system with an adversarial mindset — to help. We've designed our systems to automatically detect and filter out words or phrases that violate our policies, which prohibit users from knowingly generating content that is sexually explicit; hateful or offensive; violent, dangerous, or illegal; or divulges personal information. We also eliminate risks of exposing personally identifiable information by avoiding images with identifiable human faces. We start with more stringent filtering, and refine as we go. This work has minimized the risk, but not eliminated it, and we are continuously improving our capabilities in this area.

Imagen relies on text encoders trained on uncurated web-scale data, and thus inherits the social biases and limitations of large language models. As such, there is a risk that Imagen has encoded harmful stereotypes and representations, which guided our decision to not release Imagen for public use without further safeguards in place.

Finally, while there has been extensive work auditing image-to-text and image labeling models for forms of social bias, there has been comparatively less work on social bias evaluation methods for text-to-image models. A conceptual vocabulary around potential harms of text-to-image models and established metrics of evaluation are essential components of establishing responsible model release practices.

The Outcome

While we leave an in-depth empirical analysis of social and cultural biases to future work, our small scale internal assessments reveal several limitations that guided our decision not to release our model at this time. As is the case with generative models, Imagen may sometimes do a poor job of reflecting some parts of the data distribution, especially in the more underrepresented parts. This may further compound the social consequence of dataset bias. Imagen exhibits serious limitations when generating images depicting people. Our human evaluations found Imagen obtains significantly higher preference rates when evaluated on images that do not portray people, indicating a degradation in image fidelity. Preliminary assessment also suggests Imagen encodes several social biases and stereotypes, including an overall bias towards generating images of people with lighter skin tones and a tendency for images portraying different professions to align with Western gender stereotypes. Finally, even when we focus generations away from people, our preliminary analysis indicates Imagen encodes a range of social and cultural biases when generating images of activities, events, and objects. As a result of the AI Principles review, Google aims to make progress on several of these open challenges and limitations in future research.

Case study: A dataset for avoiding unfair gender bias

The Challenge

Researchers on the Translate[23] team recently developed a new dataset[24] for studying and preventing gender bias in machine learning in alignment with AI Principle #2,[25] "avoid unfair bias." This research explored gender translation between English and Spanish, and English and German.

The research leverages the ways different languages employ gender markers to investigate potential gender bias in translation models. Spanish is a "pro drop" language, which means subject pronouns are optional. Both Spanish and German have grammatical gender, so they mark gender on adjectives that modify people and objects. Spanish has a single possessive pronoun for his, her, and their, but English and German have separate pronouns for each. These grammatical gender differences across languages can pose a challenge for machine translation systems. This challenge is especially difficult when translating from a language without subject pronouns (such as Spanish) to one with required gendered subject pronouns (such as English).

Traditional neural machine translation (NMT) methods translate sentences one by one, but gender information often is not explicitly stated in every sentence. Seeking a novel way to address this challenge, the researchers built a new "context-aware" model that incorporates context from surrounding sentences or passages to be translated to improve gender accuracy when personal pronouns are translated.

When translating between languages with and without grammatical gender, the responsible AI challenge lies in training machine learning (ML) systems to choose the appropriate pronoun or maintain gender agreement between references throughout the content. Gender translation mistakes can be especially harmful errors, given that gender markers often convey a person's gender identity. The Translate team's new dataset was built to test the performance of this context-aware model, using gender differences across English, Spanish, and German to challenge the model to correctly translate people's genders across multiple sentences.

The Approach

The researchers applied for an AI Principles review of their dataset and proactively requested fairness testing. Reviewers and testers assessed the team's rationale for using Wikipedia biographies[26] as a source for data. The researchers chose Wikipedia biographies because the entries are well-written, geographically diverse, contain multiple sentences, and refer to subjects in the third person, using many pronouns. The reviewers and testers also looked at the researchers' strategy to prioritize equal representation of feminine and masculine identities within the dataset, while acknowledging that there were not as many biographies for non-binary people available on Wikipedia. The researchers used articles about groups that are referred to using gender-neutral "it" or "they" in English to train ML models not to incorrectly generate gendered pronouns. In addition, the reviewers and testers looked at the researchers' decision to investigate gender translation accuracy for non-Western names by sourcing Wikipedia biographies about people from 90 different nations spread across the world.

The Outcome

The result is the **Translated Wikipedia Biographies dataset**,[27] which can be used to evaluate gender bias in translation models. This dataset enables a novel method of evaluation to help reduce gender bias in machine translations. Because each instance refers to a person with a known gender, the researchers could use the dataset to compute the model accuracy of the gender-specific translations that refer to that person. This dataset provided useful performance measurements for the new context-aware models; using the dataset, researchers determined that context-aware models made 67% fewer gender translation errors than previous models that translated sentence by sentence. You can find examples of the kinds of improvements the context-aware model showed in the blog post[28] about this research.

In alignment with AI Principle #4, "Be accountable to people," the AI Principles reviewers recommended that the researchers publish a data card,[29] which is a structured document offering details about how the dataset was created and tested. With respect to AI Principle #6, "Uphold high standards of scientific excellence," the researchers decided to share the dataset publicly to support long-term improvements on ML systems focused on pronouns and gender in translation. The researchers make it clear that the dataset focuses on a specific problem related to gender bias and doesn't aim to cover all challenges of NMT, nor to be prescriptive in determining the optimal approach to address gender bias. This dataset and the research behind it aim to foster progress on this challenge across the global research community.

Case study: A more inclusive and equitable skin tone scale

The Challenge

Skin tone plays a key role in how we experience and are treated in the world, and factors into how we interact with technologies. Studies[30] show that products built using today's artificial intelligence (AI) and machine learning (ML) technologies can perpetuate unfair biases and not work well for people with darker skin tones. Computer vision (CV) is a type of AI that allows computers to "see and understand" images of people and environments, but when present-day systems aren't designed with everyone in mind, they may not "see" and "understand" people with darker skin. Building more inclusive CV systems requires being intentional — from collecting representative datasets for training and evaluation, to developing the right evaluation metrics, to building features that work for all users. In 2018, the pioneering Gender Shades[31] study demonstrated that commercial, facial-analysis APIs perform substantially worse on images of people of color and women. Gender Shades evaluated API performance across genders and skin tones, as well as intersectionality between the two. Findings from the study inspired members of the AI community to explore more inclusive CV systems and develop best practices for measuring, improving, and documenting how models and datasets represent skin tone.

The Approach

But how do we categorize the continuous spectrum of skin tones from all around the world into meaningful categories that work for evaluating and addressing fairness considerations? This is a question that Google asked over the last two years. To date, the de-facto tech industry standard[32] for categorizing skin tone has been the 6-point Fitzpatrick Scale.[33] Developed in 1975 by Harvard dermatologist Thomas Fitzpatrick, the Fitzpatrick Scale was originally designed to assess UV sensitivity of different skin types for dermatological purposes. As a result, the scale skews towards lighter tones, which tend to be more UV-sensitive. While this scale may work for dermatological use cases, relying on the Fitzpatrick Scale for ML development has resulted in unintended bias that excludes darker tones.[34]

Google Researchers reached out to Dr. Ellis Monk — an Associate Professor of Sociology at Harvard University, whose research[35] focuses on social inequalities with respect to colorism to address these biases. To develop a more representative scale, he leveraged his extensive research on skin tone and colorism in the U.S.[36] and Brazil.[37] His research focuses on how varying geographic exposure to UV radiation yields different skin tone distributions within and across ethnoracial populations. Dr. Monk consulted with experts in social psychology and social categorization, as well as underrepresented communities, to learn how they perceived the scale.

Dr. Monk's research resulted in the Monk Skin Tone (MST) Scale — a more inclusive 10-tone scale explicitly designed to represent a broader range of skin tones. The MST Scale is used by the National Institute of Health (NIH) and the University of Chicago's National Opinion Research Center,[38] and is now available[39] to the entire ML community. The MST Scale can be leveraged to evaluate datasets and ML models for better representation of people with darker skin. This year, Google's Research Center for Responsible AI and Human-Centered Technology partnered with Dr. Monk to begin to use the scale internally, with the plan to openly release the MST Scale for the larger ML community. By openly releasing the scale to the broader industry, we aim to make it possible for others to incorporate the scale into their development processes, so that ML practitioners can collectively improve this area of AI, globally.

In order to validate the Monk Skin Tone Scale for ML purposes, researchers working on skin tone fairness in Google Research launched a study within the U.S. that's now under peer review. The study's goal was to understand how well participants across diverse communities felt their own skin tone was represented within the scale, helping to identify whether ML fairness efforts with this categorization could uncover and address potential biases faced across populations. ML fairness[40] evaluations prevent common human biases from inadvertently getting reproduced by ML algorithms.

Study participants found the MST Scale to be more inclusive than the Fitzpatrick Scale and better at representing their skin tone. They also evaluated the MST Scale as being as representative of their skin tone as a 40 point beauty palette used by an industry-leading cosmetic brand, known for their inclusivity. Larger scales like these can be challenging for ML use cases, because of the difficulty of applying that many tones consistently across a wide variety of content, while maintaining statistical significance in evaluations. For example, it can become difficult for human annotators to differentiate subtle variation in skin tone in images captured in poor lighting conditions.

The Outcome

The MST Scale transforms the continuous skin tone spectrum into 10 tones that introduce enough granularity to reflect a diversity of skin tones, without increasing complexity, enabling ML training and evaluation. To improve CV systems' understanding of skin tones and improve ML fairness evaluation, we're open-sourcing the MST. Internally, we are now using the MST Scale in numerous products, including Pixel 7, Photos, and Image Search. For example, when global users make makeup-related queries in Google Images, they now see an option to further refine Search results by skin tone.

We continue to collaborate with Dr. Monk on refining the scale. We encourage ML fairness researchers, developers, and our users to offer feedback[41] on how we can improve the scale and develop our ML models responsibly, in line with Google's AI Principles.

In addition to designing and launching our education offerings, the central AI Principles operations team manages an internal hub of **self-service content** about how to put the AI Principles into practice. These include case studies and streamlined information about various services. The hub is the main entry point for requests for **ethical reviews of new AI applications** by the central AI Principles operations team.

Finally, to complement existing review procedures and frameworks, we offer a bespoke **Moral Imagination workshop** for product and research teams. The workshop enables ethical awareness and deliberation on topics relevant to AI-enabled features and product development, early on in a team's project planning process. It's been a year since we first piloted the workshop with 15 teams, and we've learned that the workshop provides a useful bridge for those who have taken existing training modules, such as our introductions to AI Principles or tech ethics, helping them to apply their knowledge to the work they face daily with their team.

We've found that creating this space can help engineers and product managers create a mental model for how, when, and why other AI ethics resources, such as technical tools or consultations on fairness, can help, and to support a team culture of ethical decision making. Through the 34 workshops delivered to date, feedback from workshop participants shows that the sessions influenced product and research strategies, and measurement of the ethical strategies' effectiveness.

Third Pillar: Structures and processes

Our third pillar of AI governance consists of the structures and processes through which we evaluate and guide our development and use of AI. We use a risk-based approach that focuses reviews in the areas they are most needed at any given time. This includes sensitive topics that change over time depending on emerging cultural or technical issues. Current examples include AI-enabled surveillance and the creation of AI-generated synthetic media often termed "fake news." With our AI Principles process, we are able to assemble a diverse set of stakeholders to ensure we consider a variety of perspectives and effectively manage risks.

Risk Assessment

This year, the central AI Principles operations team adopted an updated **risk assessment framework** (RAF) to help (1) identify, measure, and analyze ethical risks throughout the life of an AI-powered product, (2) map these risks to appropriate mitigations, and (3) develop clearer standards of acceptable risk.

This updated RAF now also draws upon the best practices of Google's cross-company Office of Compliance and Integrity and Enterprise Risk Management efforts, and is aligned with upcoming regulatory requirements in the U.S. and E.U.

With an AI Principles RAF grounded in research and a nuanced sociotechnical harms taxonomy (discussed below), we are embedding ethical risk management into Google's processes. AI Principles reviewers provide launch guidance. Depending on the product area, this may include approval, approval with recommendations, or more holistic rework depending on the risks identified within the framework.

Our AI Principles reviews and consultations prioritize the evaluation of AI's impact on humans and the environment in which a product is likely to operate. Risk is a function of the magnitude of the harm multiplied by its likelihood and frequency. Inherent risk is the amount of risk in existence absent the effect of the control environment. We ask reviewers to consider both the impact of a risk (e.g., sociotechnical harm), and the probability (likelihood, frequency) of occurrence. Inherent risk is useful in measuring and prioritizing the actual or prospective impact of a risk exposure. When thinking about the overall potential benefit and risks of harms, we also consider the harm of not launching a new technology or application, i.e., whether it might address a current harm or issues that could remain unaddressed if we did not launch it. For example, health-related AI applications, such as those that help medical specialists in diagnosing disease, might raise questions of potential fairness, privacy, or other concerns. But that must be weighed against the value of delivering solutions to a market or population that may benefit from innovations to long-standing medical challenges.

The AI Principles RAF serves as a foundational support to enable product-specific AI Principles reviews. These reviews gather detailed risk information to inform and provide context for central AI Principles reviews and other pre-launch reviews that may be necessary.

Our related harms taxonomy draws from a systematic scoping review of cross-disciplinary research, including academia, NGOs, and government documentation and guidance, which help define and describe potential harms from algorithmic systems. Developing the taxonomy involved social scientists, researchers, AI ethics program managers, and engineers, and reflects hundreds of analyses drawn from AI Principles reviews and consultations with product teams by ethics reviewers on the central operations team with multidisciplinary expertise including ethics, philosophy, human rights, user experience research, design, engineering, linguistics, law, and trust and safety.

The RAF and the harms taxonomy help to set clear expectations for product teams, scale company-level consistency, equip AI Principles reviewers with proactive risk and harm identification and response strategies, and enable decision-makers to arrive at stronger risk-based and remediation-driven outcomes.

Alignment of Principles and Reviews

In addition, we have more directly aligned our AI Principles review process with Google's general machine learning (ML) workflow for developers. This strategic approach has helped us to operationalize our capabilities for preventing, identifying, and mitigating responsible AI concerns (such as unfair bias) by advising teams as they are:

- Defining the problem that ML will solve
- Constructing or preparing data
- Building and training the model
- Evaluating its performance
- Integrating it into products and services
- Monitoring ML performance

There are some variations for specific product areas with customized AI ethics reviews, such as Cloud, Devices and Services, and Health. Cloud's Responsible AI governance process, for example, has two different types of assessments tailored to Cloud's AI development processes and enterprise contexts, from healthcare and finance to retail and entertainment.

Reviews team

This year, the **central AI Principles reviews team** conducted hundreds of ethics reviews across research and products, including proposals for future AI applications and datasets for training or evaluating ML models. In 2022, teams representing flagship products, including the Assistant, Ads, YouTube, and Search requested AI Principles reviews. Reviews spanned across large language models, datasets, community responsibility, content quality, and region-specific fairness issues in India, among other topics.

ProFair's AI Principles consultations with product teams this year have explored potential fairness issues in models and products that utilize text and image generation, person and object detection, and classification functions for teams across Google. The majority of reviews are approved for launch if the reviewers' suggestions are applied. For datasets and models, the consistent outcome is to create and publish detailed documentation of datasets and models in the form of structured transparency artifacts known as data and model cards (see the following section for details), which function like nutrition labels, providing information such as the provenance of the data (if a data card) and model performance when tested for fairness (if a model card).

Some AI launches, such as the AI Test Kitchen,[42] an online space for people outside of Google to try experimental applications utilizing LaMDA ("Language Model for Dialogue Applications"), entailed multiple reviews. **We are taking a careful approach with LaMDA to consider valid concerns about fairness, factuality, and anthropomorphization** (the tendency for humans to see human traits in inanimate objects or technologies). To date, LaMDA has gone through eleven distinct AI Principles reviews,[43] along with rigorous research and testing based on key metrics of quality, safety, and the system's ability to produce statements grounded in facts. A research paper[44] released earlier this year details the work that goes into the responsible development of LaMDA. In addition, we applied our risk framework:

Launching LaMDA for AI Test Kitchen: Applying the AI Principles Risk Assessment Framework

Intake

Identify and rate potential risks and harms inherent to generative language models and general public platforms. Assess probable social benefit.

Initial risk rating: high

Analysis

Reviewers analyze the scale and scope of a technology's potential benefits and harms, then consult with subject matter experts (SMEs) in privacy, safety, security, and ML fairness.

Risk rating confirmed

Adjustment

Reviewers and SMEs identify appropriate adjustments, mitigations, and interventions to minimize risk and sociotechnical harms.

Residual risk rating: low

Decision

Reviewers and accountable decision makers issue launch decision based on risk appetite, with social benefit weighed against residual risk.

Decision: launch

Emerging Framework for Applications We Will Not Pursue: Surveillance

To help AI Principles reviewers evaluate technology and research in the context of the four AI Applications We Will Not Pursue, which we publicly state alongside our AI Principles, we create guidelines based on trends in our growing library of cases, external research, and subject matter and community expertise.

This year we have been iterating and piloting a new Google-wide framework to inform the responsible development of applications that gather and use information within internationally accepted norms for socially beneficial uses of surveillance technologies. This emerging framework was pivotal in understanding and articulating how social benefit outweighs the potential for harm for the AI Principles review of Dynamic World,[45] a dataset that uses machine learning to enable near real-time measurement of land cover and land use, intended for positive environmental and social impact research and featured at I/O[46] this year. Dynamic World uses **Google Earth Engine**, combining satellite imagery and geospatial datasets with planetary-scale analysis capabilities.

Even with use cases like that of Dynamic World, considerations about surveillance can arise. There is often international disagreement about what counts as surveillance. Social norms are shifting at an even faster pace, sometimes leaving governments looking to the private sector as a guide for precedents to serve as guidelines. In fact, the very concept of "norms" can mean different things to different people across cultural, regional, and temporal contexts.

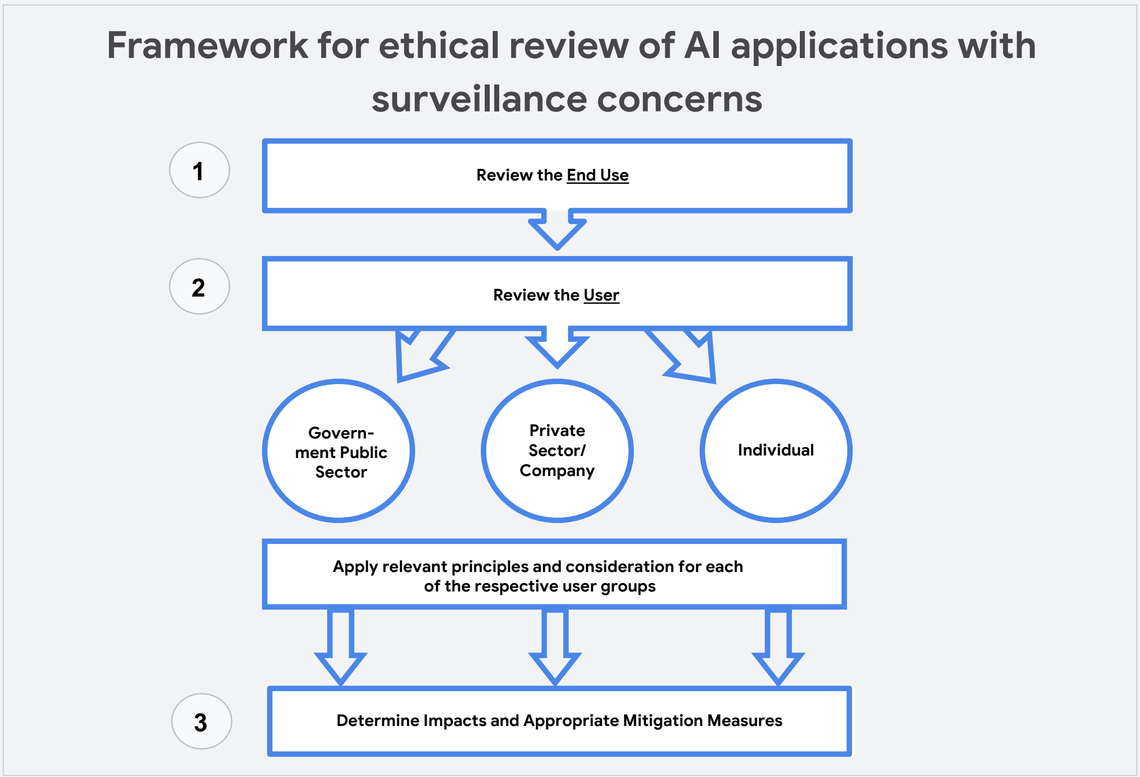
So, in 2022 we explored:

1. What a relevant international norm is, outside of the broad "big brother" definition
2. The intersections of surveillance with other AI Principles considerations, such as privacy, human rights, and downstream harm
3. When conventional applications' gathering of data might be acceptable for particular use cases

It was critical not only to work through these complex issues based on our internal understanding, but also to seek objectivity and learned expertise by engaging with external experts. Working with Google's internal Human Rights team and the external experts at Business for Social Responsibility (BSR),[47] which specializes in business and human rights, we conducted a landscape assessment of **ethical standards and norms**. One of Google's key findings was the minimum set of norms any acceptable use of surveillance technology must meet to be consistent with our privacy, human rights and ethical obligations. These include the widely known tenets of human rights: legitimacy, legality, necessity, and proportionality.

Framework for reviews

These findings helped inform our framework for assessing the Surveillance Application We Will not Pursue as part of our AI Principles reviews. The emerging guidelines consists of 3 steps:



Google's central Responsible Innovation team is piloting a framework to support AI Principles reviewers when they assess new AI projects at Google that raise potential surveillance issues

1. **Review the End Use**: Surveillance can manifest in different ways and therefore have different implications, which include categories such as targeted, mass, or indiscriminate surveillance. There are also several types of surveillance, such as tech that is used at home, in the workplace, or in public settings, and that collect video, audio, or other forms of data. Considerations vary depending on the end use, which impacts the types and potential for harm that must be assessed.

2. **Review the User**: Often, use cases involve tensions between conflicting rights

of users, such as between security and privacy. For example, during states of emergencies such as the COVID-19 pandemic, some governmental forms of surveillance might be more acceptable than during normal times for public health reasons, such as contact tracing. However, transparency and oversight are also critical factors, and explanations of AI-enabled surveillance should be easy for users to access and to understand.

3. **Mitigations**: A layer of safeguards, including guidance on how to design, use, and interpret outputs, can be implemented during the design, development or pre-launch phases. For example, model and data cards can help make potential surveillance related implications transparent. Ultimately, mitigating harmful forms of surveillance requires taking a multi-faceted approach to safeguarding users and society. The first layer of protection consists of preventative measures, such as specifying contract terms which could include feature restrictions such as mandatory face blurring. Then, ongoing ethics assessments or engaging with independent third parties to track, monitor, and report the use of technologies can be highly valuable in nascent socio-technical domains such as surveillance.

As part of our emerging framework for assessing the risk of surveillance that may violate international norms, we have customized our three-pronged **mitigation strategy**: Prevent, Mitigate, and Track. This can support a product team's efforts to address any issues before a launch.

PREVENT: Scope and Terms	MITIGATE: Partnership, Guidance, and Training	TRACK: Ongoing Review
Determining scope of use or terms of use that are rights-respecting	*Working to design, develop, and deploy technology in a rights-respecting manner*	*Reviewing effectiveness of prevention and mitigation measures over time*
• Specific contract terms which may include feature restrictions of technical limitations (such as mandatory face blurring) • Operational grievance mechanisms and/or channels to report misuse of the technology • Quality assurance (QA) procedures to verify that input data is representative and not biased or incomplete • QA procedures to identify potential unfair bias in the technology's design, testing, development, deployment, and/or outcomes • Algorithims that are explainable, and/or subject to independent assessment/audits • Restrictions on adjacent/similar use cases that are known to violate human rights standards	• Technical limitations or restrictions • Training, guidance, and direction for human operators on how to design, use, and interpret outputs • Work with users to conduct human rights due diligence before deployment • Mitigations that adequately address potential harms caused by error	• Review cycles that assess whether the technology is being used as intended and without adverse human rights impacts. May include securing feedback from the user, affected rights holders, stakeholder organizations, and other experts. • Partnering with independent organizations to track/monitor/report on the use of technology • Review of how adequately mitigations address potential harms.

Google's central Responsible Innovation team is piloting a 3-part mitigation strategy to support product teams to consider potential surveillance concerns early in the product development lifecycle

Resources, research, tools, and responsible practices

Twenty years ago, Google started using machine learning. Eleven years ago, machine learning, especially the sort known as deep learning, helped us achieve rapid progress in AI development. Our Research teams have long been at the forefront of technical AI innovation. Today, we are reinventing innovation itself by focusing on responsible practices in AI research and responsible product development.

At Google, researchers have frameworks and processes that enable them to confidently pursue and then publish research incorporating best practices for responsible AI (RAI). And our AI development teams design, and then deploy, tools that help them quickly identify and consistently remediate known problems such as unfair bias in datasets and models, so that research can be published, externalized, and incorporated into products responsibly.

Responsible AI Research

Our researchers continue to present at leading conferences around the world, including the Computer Vision and Pattern Recognition conference (**CVPR**), the Association for Computing Machinery Fairness, Accountability, and Transparency (**FAccT**) conference, the Conference on Neural Information and Processing Systems (**NeurIPS**), and more.

In 2022, we published 166 papers, on notable topics such as achieving robustness[48] and trustworthiness in large-scale ML models,[49] maintaining fairness in real-world uses of ML,[50] bringing impacted communities into AI research and development,[51] ensuring that AI is culturally competent,[52] and enabling new approaches to prototyping human-centered AI.[53]

Adversarial testing

In lockstep with Google's central operations team and processes for AI Principles reviews and governance as described in the previous section, our research teams often begin the RAI evaluation process of their technical work early in the product development lifecycle and the ML workflow, using adversarial testing — stress testing a model or product to probe it for errors and harms. To conduct adversarial testing well, we've learned to employ a variety of techniques:

- **Leverage social & cultural experts**: Google has built the ProFair team (as described earlier), to create a centralized experience in adversarial testing that incorporates global perspectives.
- **Engage testers from historically marginalized backgrounds**: Whether testing with Googlers via Google's Product Inclusion and Equity team in partnership with

volunteers from Google's Employee Resource Groups, or with a trusted external vendor, the diversity of our testers is critical to ensuring models are assessed across a wide spectrum of use-cases, scenarios, and values. We also practice thoughtful, inclusive, and equitable task design to guide humans who label data in datasets (also known as raters), in order to achieve high quality and fair evaluations.

- **Synthetic data**: We can't always scale up the number of human testers to get reliable enough results. DeepMind has made major advances[54] in using machine learning to generate synthetic data and adversarial datasets, which can preserve privacy (e.g., not require gathering personal information), offer a source of large-scale, representative data needed for model training, and provide other benefits.
- **Data quality/coverage**: One often-overlooked aspect of adversarial testing is data quality. Especially with data generated by humans, the adversarial test cases can lack diversity. This can translate into poor test coverage in which some areas are not sufficiently covered, resulting in failures that may only be discovered in production or after deployment. One of our current focus areas is researching high quality analytics to characterize adversarial prompts for large models and their outputs.

Tools

We've continued to develop and update our suite of RAI tools, which we make available to the public.

This year, we integrated the **Model Remediation library**,[55] which provides techniques for addressing bias and fairness issues in models, into our internal ML platform for automating common ML workflows across the company; it is also available publicly. The **Learning Interpretability Tool (LIT)**,[56] an updated version of the Language Interpretability Tool that we released this year, identifies how different inputs affect an ML prediction, can trace an error back to the training data, and can even measure correlations within a model. It is already being used to improve our ML models for Search. The updated LIT tool can now handle 10x the number of data sizes LIT used in the past.

Internally, researchers use tools we've developed such as **Know Your Data** (KYD),[57] which was first piloted last year. This year, we released a catalog of 70+ datasets on the tool's public-facing web site. KYD can help with responsible data analysis for ML training and evaluation sets, centering on four concepts:

1. The provenance of the data: Where did it come from?
2. The content of the data: What's in it?

3. The associations among sensitive content in the data: What is associated with what?
4. The labels of the data: Types and categories of the data

These four analytic concepts are focused on the downstream harms identified in the research literature, and are useful for coming up with recommendations such as data filtering or simply flagging results for documentation. An example data analysis of Language Models can be found in a recent paper we published on scaling language modeling with PaLM.[58]

Datasets

Our Research team is also building **high-quality datasets** to meet product needs where existing datasets are inadequate, including by leveraging third-party datasets and partnerships. Our dataset generation projects include:

- Building a new image dataset centering on a diverse representation of subjects for internal product fairness testing and development
- Creating internal synthetic datasets, built with privacy top of mind and with an emphasis on photographs that are inclusive
- Creating an externally available Wikipedia dataset to help avoid unfair gender bias

Understanding context

Our emerging best practices for responsible AI require a clear articulation of the broader societal context into which the technology may be deployed. This includes considering unfair biases, unjust impacts, and the norms and values in the real world situations in which ML models will operate. These best practices will form the basis for RAI capabilities.

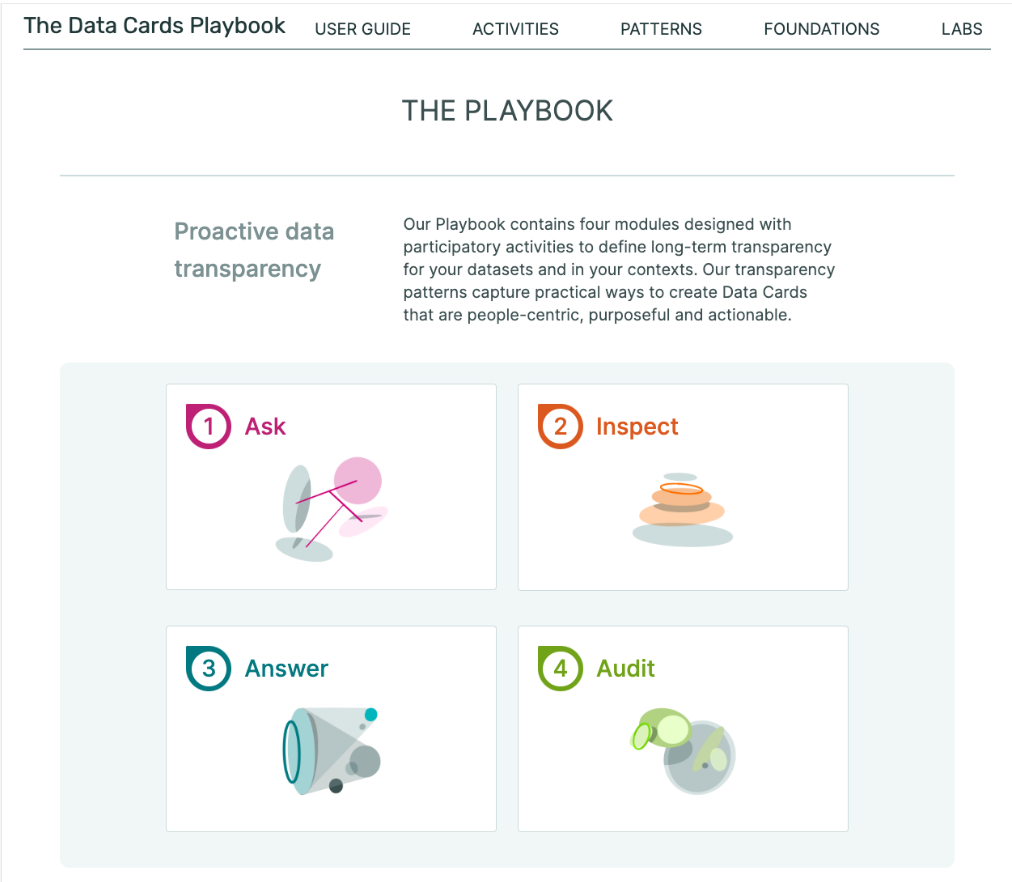
Societal knowledge is gathered and organized through various research and knowledge production efforts across our Research teams and the Office of Compliance and Integrity. The **Societal Context Understanding Tools and Solutions (SCOUTS)**[59] research initiative provides people and ML systems with the scalable, trustworthy societal context knowledge required to realize responsible and robust AI. This year, the SCOUTS **Societal Context Repository** (SCR) has shown it can help improve tools Google has shared with the world to enable responsible AI practices in important areas such as content moderation. Jigsaw,[60] Google's incubator for building technology that explores solutions to threats to open societies, used the SCR's knowledge base to supplement and balance datasets prior to model training for **Perspective API**,[61] a product used to identify toxic content in online comments. As a result, terms used for bias mitigation are fresher and have greater coverage. Additionally, the entire bias mitigation process is more scalable.[62] Perspective API is used by trusted news and information organizations around the world, including *Wired*, the *New York Times*, *El Pais*, and *Le Monde*.

Transparency artifacts

Our researchers continue to iterate on designs for new transparency artifacts to accompany our datasets, including ones that are specific to particular sectors. For example, **Healthsheets**[63] are a contextualized adaptation of the original datasheet questionnaire for health-specific applications.

These complement other structured transparency artifacts, such as **model cards**. Examples of externally available model and data cards for recent Google large models include the Parti/Imagen data card[64] and PaLM datasheet and model card.[65]

We recently launched **The Data Cards Playbook**,[66] free online guidance from the People + AI Research (PAIR) team. It's based on workshops we conducted at major AI conferences such as FAccT on how to create structured transparency artifacts to ensure data excellence: data that is clean, representative, and fit for purpose.



Google's People + AI Research team launched the Data Cards Playbook (https://pair-code.github.io/datacardsplaybook/) in 2022. The Playbook helps teams create structured transparency artifacts for datasets.

Benchmarks

Benchmarks play an important role in the RAI process for generative models that produce text outputs, images, music, and the like. Benchmarks discover areas that need fine tuning, can lead to the development of tools needed by downstream users, and establish measurable RAI baselines for comparing and understanding the strengths and weaknesses of machine learning systems. This year we've updated our **RAI language benchmarking capabilities** to include measurements related to toxic language, unfair discrimination, and expressions of stereotypes or other forms of marginalization. We also developed benchmarks for text-to-image generation tasks to include sensitivity to changes in context and style of images.

Responsible Generation

Responsible generation refers to methods that help control the outputs of generative models so that they better meet some responsibility goals. For example, we may not want a chatbot to reply with toxic language, even if coaxed.

There are a number of recent advances that are very promising for responsible generation that we are exploring for applications of new generative models like PaLM:

- There has been an explosion in research into how to prompt language-driven models such as DALL-E, Imagen, GPT3, and PaLM to help control or improve the models' output. For example, Google's discovery of **"chain of thought"**[67] prompting has illustrated how including the reasoning steps can significantly improve language models responses. Methods to search for prompts have also been used to identify what kinds of prompts lead to unintended model behavior.[68]
- A particularly exciting and relatively new approach to control, that goes beyond using prompt-text, is called **parameter efficient tuning**. It has recently been shown that small datasets can be used to tune a special input to a model that results in significantly better control.[69]

Another significant branch of work to control language models focuses on fine-grained control of the output, instead of the inputs. For example, methods have been developed to reduce the likelihood of toxic outputs[70] and more recently to mitigate unintended biases.[71]

Technical Guidance: Responsible AI Product Maturity Model Assessment

To help scale responsible AI emerging best practices and to monitor and control model performance over time, our Responsible AI and Human Centered Technology (RAI-HCT) team, which focuses on the technical realization of the AI Principles, developed and internally deployed the **Product Maturity Model Assessment (PMMA)** in June.[72]

The PMMA helps teams across Google who leverage machine learning in their products to:

- Understand the RAI domain and its emerging best practices and tools
- Measure the maturity of their ML models and use cases
- Proactively improve the maturity of their ML models (i.e., achieve desired RAI outcomes) and use cases over time

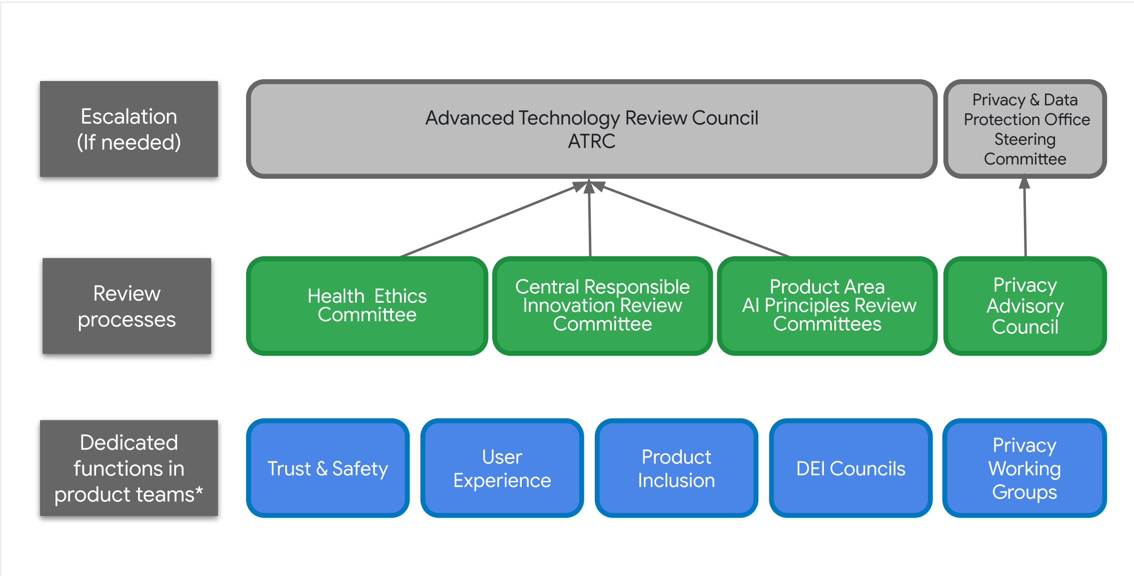
The PMMA centers on an extensive questionnaire with technical and application-specific questions based on the current responsible practices in the field of AI research and development. These questions were sourced from the latest research in the emerging academic field of responsible AI, both internal and external to Google, along with emerging RAI best practices currently being implemented by Google product teams. A cross-functional team of research scientists, engineers, product managers, user experience specialists, technical writers, and program managers generate and refresh these PMMA questions to reflect new advances in technical knowledge.

The PMMA is designed to gauge adoption of state-of-the-art RAI practices that reflect the AI Principles. It maps survey results against a maturity model framework that is prescriptive in nature, with clear courses of action provided to help teams improve their ML models from one maturity level to the next one. Teams access their PMMA results via interactive dashboards and visualization tools, and can apply this critical input when adjusting their ML models and more closely aligning them with the AI Principles.

In addition to understanding where their ML models need to improve, product teams that have taken the PMMA consistently ask for concrete steps on how to start and maintain responsible practices over time. To meet that need and enable RAI across Google, we are building out a **library of relevant technical guidance** documents and tooling. We are also creating a **centralized technical infrastructure** that will eventually automate most of the RAI technical tasks that underlie our PMMA questions, allowing product teams to conduct these tasks in a manner that is repeatable, accurate, and built for scale.

Product Impact

There are three tiers of complementary, company-wide governance structures in place to put the AI Principles into practice for all teams building products across Google.



There are dedicated functions that support responsible AI practices embedded within Google's product teams. Google further operationalizes these practices across the company via a three-tiered internal AI Principles Ecosystem.

The first tier exists within product teams themselves. Product teams include dedicated user experience (UX), privacy, and trust and safety (T&S) experts, providing deep functional expertise consistent with the AI Principles.

The second tier is the set of dedicated review bodies and expert teams. This includes the central AI Principles team in the Office of Compliance and Integrity, discussed earlier in this report. It also includes specialized reviewers with expertise in specific areas, such as Cloud, with expertise in enterprise AI offerings, and Devices and Services, with expertise on how AI is used responsibly within hardware. There is also a Health Ethics Committee that reviews health and medical-related AI and ML research and development across the company. All Google employees are encouraged to engage with the AI Principles review processes throughout the project development lifecycle. In addition, the Privacy Advisory Council (PAC) reviews all projects for potential privacy concerns, including (but not exclusively) issues related to AI.

The third tier of our AI governance structure is the Advanced Technology Review Council (ATRC), a rotating committee of senior product, research, and business executives. This Council represents a broad cross-section of Google. The ATRC addresses complex cases that could affect multiple products, or set a precedent because the technology is so new. The Council also makes decisions on urgent escalations, and establishes company-wide policies impacting multiple product areas. This involves making challenging decisions that require deeply considering the trade offs between ethical risks of certain new applications and potential business opportunities, prioritizing social benefit. For example, this year, the ATRC completed two reviews of LaMDA, deciding to prioritize rigorous, **responsible generative AI development**, testing, and guardrails above speeding the technology to market.

Directly or indirectly, our responsible AI research, training, and tools were important contributors to ensuring that the AI Principles are comprehensively reflected within the wide spectrum of product innovations we announced throughout the year.

For Example:

AIP #1: Be socially beneficial. This principle helps us think through how the overall benefits of AI exceed risks, including potential privacy or other concerns. Because we are an information company, we strive to use advanced AI to help people who use our products and services make high-quality content available.

In 2022, the **spread of misinformation** has been a growing global challenge. Because of the amount of misinformation that's created and the speed with which it spreads, it seems like a natural for an AI-based solution…or at least a mitigation. But, for that we would need a large training dataset of expertly-curated misinformation. To help address this — as we've outlined on YouTube's blog[73] — we're continuously training **YouTube**'s system on new data that's been collected responsibly and with clear user policies for gathering informed consent. YouTube is looking to leverage an even more targeted mix of classifiers, keywords in additional languages, and information from regional analysts to identify narratives our main classifiers don't catch. Over time, this will make YouTube faster and more accurate at catching viral misinformation narratives. YouTube is also working on ways to update models more often in order to catch hyperlocal misinformation, with the capability to support local languages.

We also look for opportunities to support humanitarian uses of AI. For example, to help refugees and veterans of the war in Ukraine, we updated **Look to Speak**,[74] an Android app that allows people to use their eyes to select pre-written phrases and have them spoken aloud in 18 languages, to include Ukrainian.

As a broad approach to social benefit, this year we launched AI for the Global Goals,[75] an initiative to bring together research, technology, and funding to accelerate progress on the United Nations Sustainable Development Goals (SDGs). This commitment will include **$25 million to support NGOs and social enterprises working with AI to accelerate progress towards the SDGs**. Based on what we've learned so far, we believe that with the AI capabilities and financial support we will provide, grantees can **cut in half the time or cost to achieve their goals**. We'll also provide Google.org Fellowships, where teams of Google employees work alongside organizations for up to six months.

AIP #2: Avoid creating or reinforcing unfair bias. Unfair bias exists in the world, and datasets and models could reflect these, including unintentionally. This principle helps us consider from the earliest ideation stage how to harness AI so we can truly build for everyone, and for cultures around the world.

We continue to create new programs to address emerging challenges in fairness. For example, the recently launched **Equitable Automated Speech Recognition program** at Google aims to identify and eradicate unfair biases in voice technologies.[76]

As a company with the mission of making the world's information universally accessible and useful, a more global approach to enabling communication for more people across the world — no matter their language or abilities — was a clear focus of some of our biggest product announcements.

The **Project Relate** app, which helps improve communications for those with impaired speech, has expanded globally and is now available in the U.S., Canada, Australia, New Zealand, India, and Ghana.[77] We also expanded **Project Euphonia**, a research initiative that works with community organizations and people with speech impairments to create more inclusive speech recognition models. This year we added French, Hindi, Japanese and Spanish.[78]

We also launched the **1,000 Language Initiative**[79] to support the 1,000 most-spoken languages in the world, using our universal speech model trained on more than 400 languages — the largest language coverage in a speech model to date. We also used ML to add 24 new languages to Translate,[80] a technical milestone for **Google Translate**, for these are the first languages we've added using Zero-Shot Machine Translation, which can be trained using texts only in the new language, without translations from that language. This will enable Google to accelerate its addition of new languages, especially ones with relatively few translated texts available. This will help give voice to the world's less-heard populations.

AIP #3: Be built and tested for safety. We design our AI systems with strong security practices, to avoid user harms — including those from unintended uses or results. For example, starting in March of this year, Chrome rolled out a new ML model that identifies 2.5 times more potentially malicious sites and phishing attacks as the previous model — resulting in a safer and more secure web.[81]

AIP #4: Be accountable to people. Using our emerging practices for designing human-centered AI systems as captured in the public-facing **People + AI Research Guidebook** — which we first released in 2019 and update regularly — we design AI systems with humans in the loop, clear opportunities for user feedback, relevant explanations, and appeal.[82] This year, we expanded upon our offerings in UX by launching a Google-wide Explainability course, which we also presented externally.[83]

AIP#5: Incorporate privacy by design principles. To safeguard users' privacy, we apply Google's cross-company privacy principles, which pre-date the AI Principles. This year, we announced several new uses of on-device models to keep information private, one of our main proactive practices across Google and in alignment with our AI Principles. These ranged from more convenient transcription options in Pixel 7's Recorder feature[84] to Web browsing with minimal interruption via a new system for more relevant permission prompts in Chrome.[85]

We continue to use other privacy design strategies, too. For example, for a pilot to update **Google Maps**[86] with updated speed limit information, Google requests photos from trusted third-party imagery partners that already gather roadway imagery to improve delivery routes. Those photos are of specific stretches of road that also include a speed limit sign. If the partner has this photo available, we then use a combination of AI and help from our operations team to identify the sign in the image, and extract the new speed limit information to add to Google Maps. We reference only images taken on public roads, and partners are required to blur identifying information such as faces and license plates. For an extra layer of privacy, we blur the photo again when we receive it and delete the photo after we use it to update the map.

And at the I/O conference, we announced that when user information is sent to Google's servers, more of it will be anonymized through techniques including the use of differential privacy and edge computing.[87]

AIP #6: Uphold high standards of scientific excellence. Innovation in AI, or in any technical field, is rooted in the scientific method, requiring intellectual rigor, multidisciplinary collaboration, and knowledge sharing.

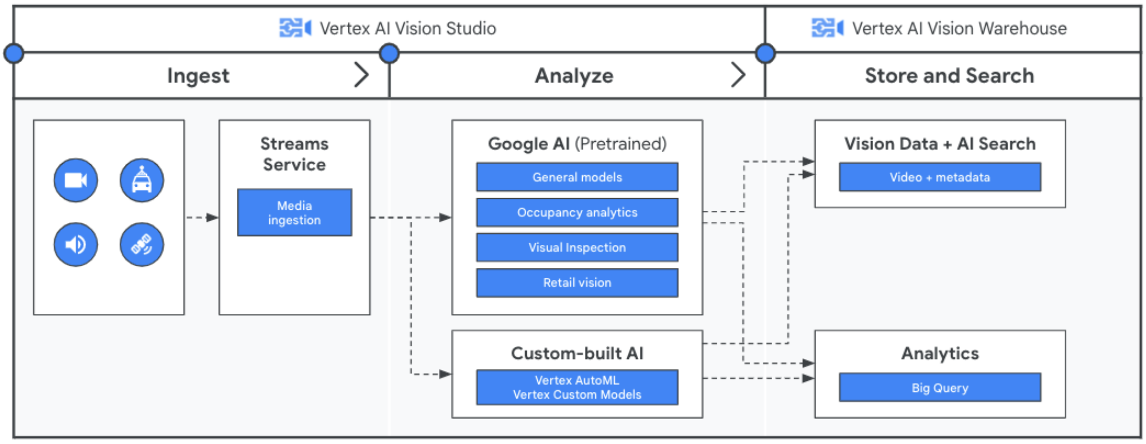
Google and our parent company Alphabet have continued to be important contributors of scientific research in papers, with topics on AI that range from applying neural networks to restore ancient texts to integrating deep learning for disease detection, as evidenced by our place in the top 5 of global corporate research in **the 2022 *Nature* index**.[88]

To make our responsible AI research more accessible and findable, in early 2022, we curated a dedicated external collection of more than 200 research papers focused on the topic of Responsible AI.[89]

AIP #7: Be made available for uses that accord with these principles. As we develop AI, we utilize this principle to determine how to limit any potentially harmful deployments, including how the AI could be adapted. We also consider our role in providing general-purpose tools, as well as developing custom AI solutions, or integrating tools for our enterprise or other customers.

Our responsible AI approach prioritizes in-depth evaluations during development and guidance for our enterprise customers once implemented so they can safely build and deploy applications using Google Cloud, accounting for their unique social and organizational contexts. For example, to align Cloud's **Vertex AI Vision**[90] with the AI Principles, the Cloud Responsible AI team conducted an evaluation and incorporated mitigations for risk concerns and education opportunities, including:

- Evaluating and testing for unfair bias during development
- Developing product features to enhance privacy and limit personal identification
- Increasing product documentation and transparency to support customers' responsible use



Google Cloud's Vertex AI Vision was evaluated by Cloud's Responsible AI review team for unfair bias, privacy and transparency before launch.

Supporting global dialogue, standards, and policy

Practicing responsible AI research and development requires a collective approach, so it is vital to share learnings and receive feedback on our progress from the world's larger AI ecosystem.

We continue to develop publicly available content that explains how the AI in our core products and services work, in the tradition of *How Search Works*[91] and *How YouTube Works*.[92] We realize raising awareness is the first step in supporting international conversations about responsible AI. So we share our learnings and emerging best practices via free, online educational content and programs.

This year, we increased our AI Principles online courses and other educational programs for a wide-range of external audiences. These include:

- **"Discover AI in Daily Life,"**[93] a course designed with middle and high school students in mind from Applied Digital Skills,[94] Google's free, online, video-based curriculum (and part of the larger Grow with Google[95] initiative) explains how AI is built, how it helps people every day, and the potential challenges AI faces and poses. Building AI literacy is an important component of accountability, as it helps prepare people to participate in civic discussion and understand explanations in products.[96] The lesson has over 68,000 unique users to date.
- A new external training, **"Responsible AI: Applying AI Principles with Google Cloud"**[97] is designed to provide a framework for any organization interested in operationalizing responsible AI practices.
- To make our research more accessible and useful, we launched **an interactive scientific article**[98] where people can play with our ML models — getting results in real time, in their web browser, with no setup required.

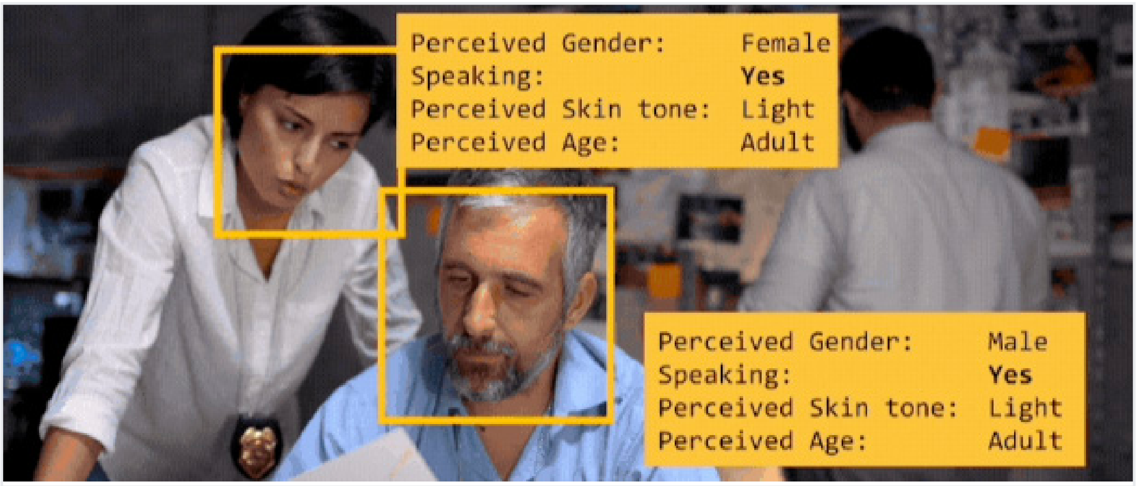
We continue to support academic and research partnerships and fellowships to increase diversity, equity, and inclusion in the field of computer science. Highlights from this year include:

- Serving as a founding partner of the inaugural **African Master of Machine Intelligence (AMMI)** program. Many AMMI graduates have continued their studies or taken positions in industry,[99] including at our Accra Research Center where we offer an AI residency program. We've had three cohorts of AI residents to date.
- Launching Google's first-ever **PhD Fellowships** for students attending Latin American universities. Selected Fellows receive $15,000 of funding for up to three years, and mentorship from a Google researcher to support their career

development in Computer Science Research.

- Expanding the **Responsible Innovation Fellowship**[100] program from an internal Google fellowship program to a 5-week external program that equips students with the knowledge and skills they need to enter the field of AI ethics. Students from currently underrepresented backgrounds — including students from U.S. Minority Serving Institutions, such as Historically Black Colleges & Universities, Hispanic-Serving Institutions, and Historically Women's Colleges — are encouraged to apply. The inaugural cohort engaged twenty students.[101]
- Announcing a $20 million commitment to **expand CS education** access to more than 11 million students across the U.S. This brings our total commitment to CS education to more than $240 million since 2004.[102]
- Hosting the largest cohort of Google's **CS Research Mentorship Program** (CSRMP),[103] to date, with 300 students in 2022. CSRMP aims to increase the diversity of PhD graduates in computing-related fields and ensure the broader community of CS researchers includes the experiences, perspectives, and concerns of people worldwide. Since 2018, CSRMP has hosted more than 730 students across more than 230 institutions.[104]
- Developing a **business school case study** on operationalizing AI Principles targeted for future technology industry decision-makers, in partnership with the *California Management Review*.[105] The case is now taught in multiple courses on AI and ethics at the University of California, Berkeley, Haas School of Business.
- Investing in the launch of **INSAIT**,[106] the Institute for Computer Science, Artificial Intelligence and Technology, a new AI and computer science research institute in Bulgaria, backed by the Bulgarian government with an endowment fund of nearly $100 million.[107] Its research will include machine learning, quantum computing, information security, robotics, and more. Google is investing $3 million over the next three years to provide INSAIT with cloud computing resources and access to Tensor Processing Unit Research Cloud,[108] a specialized infrastructure for running high-performance machine learning models.
- Collaborating on the **See It, Be It: What Families are Seeing on TV** study and report[109] to analyze trends in the screen and speaking time of characters based on their perceived gender, skin tone, and age, in scripted television over the last 12 years. The research was led by the Geena Davis Institute on Gender in Media,[110] in partnership with Google Research as the technology provider, and the Signal Analysis and Interpretation Laboratory[111] at the University of Southern California as the academic advisor. Together, we applied a new skin tone classifier based on the MST Monk Skin Tone Scale to study representation patterns in a socio-technical

system, to examine representation in Nielsen's top 10 scripted U.S. TV shows for each season from 2010 to 2021.



Google Research provided the technology, including a new skin tone classifier based on the MST Monk Skin Tone Scale, to a new study examining representation across gender, skin tone and age in scripted TV over the last 12 years,

We also take a holistic approach to prepare more people for future careers in the AI industry around the world, even if they are not formally studying or have never studied computer science or programming:

- In the U.K., we announced a **partnership with Girlguiding**[112] that will provide nearly 400,000 girls training in concepts such as coding and algorithms, with new activities co-created by Google's women engineers. The new activities will form part of Girlguiding's national program within the Skills for my Future theme. These span all four Girlguiding sections (age groups) and have been created to be completed offline to ensure they are accessible to all girls.
- In June at APAC, we launched[113] the Japan Reskilling Consortium, a **collaboration between business, governments, and the nonprofit sector** that provides skills training in areas such as artificial intelligence and digital marketing. It also provides a job-matching service to help trainees find work opportunities. The consortium already offers more than 300 training programs with more than 90 partners.
- In the U.S., we announced[114] that we will be joining Ford Motor Company as a **founding member of Michigan Central**.[115] Michigan Central is a new innovation hub where companies, government, and community stakeholders will focus on the future of mobility — both in terms of economic opportunity and transportation solutions — in Detroit and beyond. Specifically, we're offering our Cloud infrastructure, AI and ML capabilities, and data and analytics tools to Michigan Central to be used on projects and research for future mobility solutions.

Alongside policymakers and standards organizations, we recognize AI carries risks as well as opening up exciting possibilities. We are optimistic about the potential for standards to continue to move the industry forward in a responsible way. So we remain actively engaged in key conversations with international forums such as the Organisation for Economic Co-operation and Development (**OECD**), the International Standardization Organization (**ISO**), and national standards bodies such as **NIST** in the U.S.

For example, in collaboration with the private and public sectors — including dialogue with the EU, ISO, and others — NIST is developing a voluntary-use framework to help organizations better manage risks associated with AI. We filed a **Google response to NIST's request for feedback**[116] on an initial framework draft. We expressed our support for the initiative and shared recommendations, including to clarify the roles of different stakeholders in the AI value chain, expand on the distinction between fairness and unfair bias, and determine how the NIST framework can be integrated with other standards and frameworks.

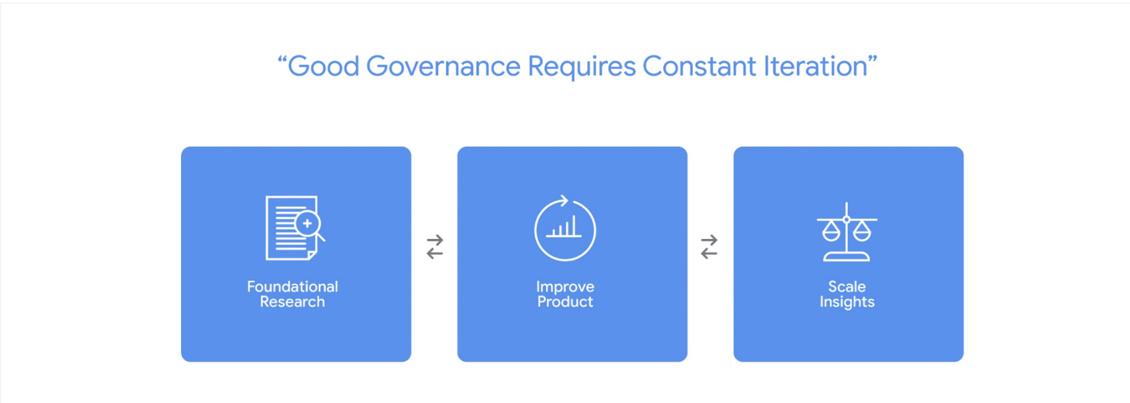
Through these relationships, we work to help coordinate efforts to drive the development of consensus, multi-stakeholder standards for AI and emerging tech systems, and common benchmarks for AI evaluation. We will continue to contribute to the **public consultation on emerging legislation like the EU's AI Act,** and share resources with policy leaders with our **Machine Learning for Policy Leaders workshop**, which has trained 600 stakeholders around the world since its launch in 2020. We continue to partner with other external organizations, including the Future of Humanity Institute, University of Oxford, and The Centre for Internet and Society, among many other engagements.

Conclusion

AI will always reflect human values. For AI to live up to its full potential, the complex intersection of those values with this new technology will require multidisciplinary thinking across computational and social sciences — including foundational research, multimodal content understanding, a responsible AI lens, and more. It is essential that this truly collective approach involves many international and interdisciplinary contexts. That is why we are committed to a collaborative, cross-disciplinary approach to building helpful AI for everyone.

This year, to continue to build on this strategy, we supported the *Economist Impact*'s efforts to publish whitepapers on the economic impact of AI and how governments can encourage AI adoption in the Middle East and North Africa[117] and Latin America.[118] We shared insights from these papers at respective events co-hosted with the UAE AI Ministry and the Council of the Americas, and continue to engage with key policymakers and stakeholders in these regions.

Internally, we continue to engage external advisors with our socio-technical AI research that informs AI research and product design. For example, to identify potential fairness considerations, the Photos and Assistant product teams, and researchers working on **generative AI including Imagen, Parti, and the AI Test Kitchen, have participated in the Equitable AI Research Roundtable** (EARR).[119] This program offers the opportunity to identify potential fairness considerations with a group of experts from the Othering and Belonging Institute at UC Berkeley; PolicyLink; Ed Trust West; University of Texas, Austin; and Emory University School of Law early in the product development lifecycle.



Google's AI Principles governance strategy is to constantly iterate and improve – which also reflects the company's overall responsible innovation strategy

We still have a lot to learn — and will continue learning — given the dynamic and evolving nature of technology and society. The act of innovation is rooted not only in coming up with Big Ideas but also in constant iteration, including within our own AI governance and operations, processes, and tools.

Innovation also means reaching out to users and partners across the technology industry, geographies, academic disciplines, cultures, and governments, listening to and analyzing their needs over time. This helps us design and test, learn from missteps, adjust, and improve. Listening helps us proactively design new AI solutions as new challenges emerge — and respond to feedback, and the inevitable criticisms, with respect and a commitment and willingness to change.

Because AI helps us address the world's most pressing large-scale problems and helps to shape the world's future, it's more important than ever that we get this right, together.

End Notes

1. https://blog.google/technology/ai/join-us-in-the-ai-test-kitchen/
2. https://blog.google/products/translate/24-new-languages/
3. https://blog.google/technology/health/advancing-genomics-better-understand-and-treat-disease/
4. https://blog.google/technology/health/check-up-ai-developments-2022/
5. https://www.wired.com/story/hurricane-ian-destroyed-homes-google-algorithms-sent-money/
6. https://www.iso.org/committee/6794475.html
7. https://www.nist.gov/itl/ai-risk-management-framework
8. https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework
9. https://www.mas.gov.sg/schemes-and-initiatives/veritas
10. https://www.niti.gov.in/sites/default/files/2022-11/Ai_for_All_2022_02112022_0.pdf
11. https://www.justice.gc.ca/eng/csj-sjc/pl/charter-charte/c27_1.html#:~:text=In%20addition%2C%20the%20Artificial%20Intelligence,with%20other%20government%20entities%20specified
12. https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=1853928&filename=PL-21-2020
13. https://www.msit.go.kr/bbs/view.do?sCode=eng&mId=10&mPid=9&bbsSeqNo=46&nttSeqNo=9
14. https://www.whitehouse.gov/ostp/news-updates/2021/10/22/icymi-wired-opinion-americans-need-a-bill-of-rights-for-an-ai-powered-world/
15. https://www.gov.uk/guidance/data-ethics-and-ai-guidance-landscape
16. https://www.gov.il/en/departments/news/most-news20221117
17. https://www.ft.com/content/3467659a-386d-11ea-ac3c-f68c10993b04
18. https://blog.google/technology/ai/crossword-puzzle-big-purpose/
19. https://blog.google/technology/ai/an-update-on-our-work-in-responsible-innovation/#:~:text=moral%20imagination%20workshop
20. https://blog.google/inside-google/googlers/meet-3-women-who-test-google-products-fairness/
21. https://parti.research.google/
22. https://imagen.research.google/

23. https://translate.google.com/
24. https://storage.googleapis.com/gresearch/translate-gender-challenge-sets/Readme.html
25. https://ai.google/principles/
26. https://en.wikipedia.org/wiki/Wikipedia:Biographies_of_living_persons
27. https://storage.googleapis.com/gresearch/translate-gender-challenge-sets/Readme.html
28. https://ai.googleblog.com/2021/06/a-dataset-for-studying-gender-bias-in.html
29. https://storage.googleapis.com/gresearch/translate-gender-challenge-sets/Data%20Card.pdf
30. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf
31. http://gendershades.org/
32. https://openaccess.thecvf.com/content/CVPR2021W/RCV/papers/Hazirbas_Casual_Conversations_A_Dataset_for_Measuring_Fairness_in_AI_CVPRW_2021_paper.pdf
33. https://www.ncbi.nlm.nih.gov/books/NBK481857/table/chapter6.t1/
34. https://ieeexplore.ieee.org/abstract/document/9590512
35. https://direct.mit.edu/daed/article/150/2/76/98313/The-Unceasing-Significance-of-Colorism-Skin-Tone
36. https://www.journals.uchicago.edu/doi/abs/10.1086/682162
37. https://scholar.harvard.edu/files/monk/files/monk_-_the_consequences_of_race_and_color_in_brazil_-_sp.pdf
38. https://www.norc.org/Pages/default.aspx
39. https://skintone.google/get-started
40. https://developers.google.com/machine-learning/crash-course/fairness/video-lecture
41. https://google.qualtrics.com/jfe/form/SV_eFJF7qguvcWvdFs
42. https://blog.google/technology/ai/join-us-in-the-ai-test-kitchen/l
43. https://ai.googleblog.com/2022/01/lamda-towards-safe-grounded-and-high.html
44. https://arxiv.org/abs/2201.08239
45. https://dynamicworld.app/explore/
46. https://io.google/2022/program/385b422e-3a08-4372-8ed3-4e79bafb779a/
47. https://www.bsr.org/en/
48. https://arxiv.org/abs/2110.07858

49. https://arxiv.org/pdf/2205.05256.pdf
50. https://arxiv.org/pdf/2202.01034.pdf
51. https://dl.acm.org/doi/abs/10.1145/3491102.3517716
52. https://ai-cultures.github.io/
53. https://dl.acm.org/doi/abs/10.1145/3491101.3503564
54. https://www.deepmind.com/publications/red-teaming-language-models-with-language-models
55. https://www.tensorflow.org/responsible_ai/model_remediation
56. https://pair-code.github.io/lit/
57. https://knowyourdata.withgoogle.com/
58. https://arxiv.org/abs/2204.02311
59. https://sites.research.google/scouts/
60. https://jigsaw.google.com/
61. https://perspectiveapi.com/
62. https://medium.com/jigsaw/scaling-machine-learning-fairness-with-societal-context-be73d4ad38e2
63. https://arxiv.org/abs/2202.13028
64. https://github.com/google-research/parti/blob/main/data_cards/fit400m_data_card.pdf
65. https://arxiv.org/abs/2204.02311
66. https://pair-code.github.io/datacardsplaybook/
67. https://arxiv.org/abs/2201.11903
68. https://aclanthology.org/D19-1221.pdf
69. https://arxiv.org/abs/2205.05638
70. https://arxiv.org/abs/2009.06367
71. https://aclanthology.org/2022.acl-long.72.pdf
72. https://blog.google/technology/ai/an-update-on-our-work-in-responsible-innovation/#:~:text=an%20internal%20tool%20to%20help%20teams%20assess%20how%20ml%20models%20were%20developed
73. https://blog.youtube/inside-youtube/inside-responsibility-whats-next-on-our-misinfo-efforts/
74. https://blog.google/outreach-initiatives/accessibility/look-to-speak-launches-in-ukraine/

75. https://globalgoals.withgoogle.com/
76. https://medium.com/people-ai-research/q-a-courtney-heldreth-and-michal-lahav-on-addressing-inequitable-speech-recognition-it-takes-a-a2d65b1b7744
77. https://blog.google/outreach-initiatives/accessibility/project-relate/
78. https://sites.research.google/euphonia/about/
79. https://blog.google/technology/ai/ways-ai-is-scaling-helpful/
80. https://ai.googleblog.com/2022/05/24-new-languages-google-translate.html
81. https://blog.google/products/chrome/building-a-more-helpful-browser-with-machine-learning/
82. https://pair.withgoogle.com/guidebook/
83. https://blog.google/technology/ai/helping-people-understand-ai/
84. https://blog.google/products/pixel/feature-drop-december-2022/
85. https://blog.google/products/chrome/building-a-more-helpful-browser-with-machine-learning/
86. https://blog.google/products/maps/how-ai-and-imagery-build-self-updating-map/
87. https://blog.google/technology/safety-security/how-we-make-every-day-safer-with-google/
88. https://www.nature.com/nature-index/institution-outputs/articles/all/global/United%20States%20of%20America%20%28USA%29/Alphabet%20Inc./corporate
89. https://research.google/pubs/?collection=responsible-ai
90. https://cloud.devsite.corp.google.com/vertex-ai-vision
91. https://www.google.com/search/howsearchworks/
92. https://www.youtube.com/howyoutubeworks/
93. http://g.co/DiscoverAI
94. https://applieddigitalskills.withgoogle.com/s/en/home?utm_source=keyword&utm_medium=blog&utm_campaign=20220211-AI-Blog-2022--all-all-&src=keyword-blog-20220211-AI-Blog-2022--all-all-
95. https://grow.google/
96. https://blog.google/technology/ai/helping-people-understand-ai/
97. https://www.cloudskillsboost.google/course_templates/388
98. https://google-research.github.io/proteinfer/
99. https://aimsammi.org/blog-post/aims-launches-african-masters-in-machine-intelligence-in-

kigali-rwanda/

100. https://research.google/outreach/phd-fellowship/

101. https://blog.google/outreach-initiatives/education/a-look-at-the-responsible-innovation-fellowship/

102. https://blog.google/outreach-initiatives/education/expand-cs-ed-access/

103. http://g.co/csrmp

104. https://blog.google/technology/research/mentorship-inspires-deyrel-diaz-and-future-researchers/

105. https://store.hbr.org/product/responsible-a-i-tackling-tech-s-largest-corporate-governance-challenges/b6021?sku=B6021-PDF-ENG

106. https://insait.ai/

107. https://blog.google/technology/ai/investing-in-eastern-europes-ai-future/

108. https://sites.research.google/trc/about/

109. https://seejane.org/research-informs-empowers/see-it-be-it-what-families-are-watching-on-tv/

110. https://seejane.org/

111. http://sail.usc.edu/

112. https://blog.google/around-the-globe/google-europe/united-kingdom/girlguiding-and-google-technology-is-for-everyone/

113. https://www.blog.google/around-the-globe/google-asia/digital-skills-japan/

114. https://blog.google/outreach-initiatives/grow-with-google/michigan-central-ford/

115. https://michigancentral.com/

116. https://www.nist.gov/itl/ai-risk-management-framework

117. https://impact.economist.com/perspectives/technology-innovation/pushing-forward-future-ai-middle-east-and-north-africa

118. https://impact.economist.com/perspectives/technology-innovation/seizing-opportunity-future-ai-latin-america

119. https://blog.google/technology/ai/update-our-progress-responsible-ai-innovation/#:~:text=equitable%20ai%20research%20roundtables%20(earr)%2C

This page is intentionally left blank.

Google