Google

AI Principles
2020 Progress update

Table of contents

Overview

Google's AI Principles were published in June 2018 as a charter to guide how we develop AI responsibly and the types of applications we will pursue. This report highlights recent progress in AI Principles implementation across Google, including technical tools, educational programs and governance processes.

Of particular note in 2020, the AI Principles have supported our ongoing work to address systemic racial bias, through learning from a diverse set of voices internally and externally, new technical methods to identify and address unfair bias, and careful review. The AI Principles were also part of Google's Rapid Response review process for COVID-19 related research.

One of the most challenging aspects of operationalizing the Google AI Principles has been balancing the requirements and conditions of different Principles. For example, how should the collection of personal data be minimized, while also ensuring that products don't have unfair outcomes for specific groups of people? Should an AI application that offers significant accessibility benefits be open-sourced, if it could also present a high risk of abuse outside its intended purpose? This report reflects our learnings in making sure that any such trade-offs increase fairness and equity overall, and builds on progress made since our 1 year update, which is available to view at http://bit.ly/2019AIP-Report.

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.

Culture, education, and participation

Implementing our AI Principles starts with equipping all Googlers with the ability to understand how the Principles apply to their own work. In addition to dedicated training for specific roles, like product managers or machine learning specialists, we have developed educational materials and courses on AI ethics for all our employees, and for others outside Google interested in learning about responsible AI.

Since 2019, we've expanded and refined many of the courses offered, ranging from high-level ethics courses and ML basics to in-depth training on how to address AI Principles issues internally. This includes an online, more accessible version of our internal in-person Technology Ethics Training. The course's lessons and self-study videos teach employees about the philosophy and core tenets of technology ethics, and how to assess whether an AI application might cause physical, psychological, social, or environmental harm.

Our AI Principles in Practice training, launched in late 2019, builds on these ethical tenets to educate Googlers about the AI Principles and our internal governance practices. The course provides a foundational understanding of Google's AI Principles, and equips Googlers in a wide variety of roles with the tools to recognize, resolve, and escalate AI Principles issues appropriately. More than 2,000 Googlers—including engineers, researchers, product and program managers, and policy and legal teams—completed an in-person pilot of this training before it was curtailed by the closure of offices due to the pandemic. In Spring 2020, we launched an hour-long self-study version online, available to all Google employees. To date, AI Principles in Practice training has been taken by over 10,000 employees.

In addition to internal training, we also offer public courses for the broader AI community. Our publicly available ML Crash Course, for example, provides a structured series of 25 lessons on everything from the basics of machine learning to how to implement ML applications responsibly. More than 78,000 people outside of Google have taken the Crash Course module on ML Fairness.

Technical progress

Since our last report, we have made substantial progress in publishing research and developing tools and techniques to help developers incorporate responsible practices into their work.

The Appendix provides details on some of the more than 200 research papers and articles on AI ethics, fairness, explainability, privacy, and security that we have published since our last report. Among the research published we have documented new techniques to improve fairness in machine learning at scale[1], to improve fairness without demographic information on users[2], and to incorporate ethical principles directly into machine learning models.[3] We also developed design principles for interpretable machine learning systems.[4]

We also continue to release datasets to support the wider research and developer ecosystem. To help researchers address misinformation, for example, we released a large dataset of 3,000+ visual deepfakes to help improve deepfake detection techniques. The dataset was incorporated into Technical University of Munich and the University Federico II of Naples' FaceForensics benchmark.[5]

In addition to basic research, we've continued to develop new tools and infrastructure to support internal and external AI developers. In late 2019 we launched Fairness Indicators, a suite of fairness testing tools that allow model developers to understand their data and potential biases in their models. Internally, Fairness indicators are now used to inform product teams such as YouTube, Jigsaw and Cloud.

Where possible, we strive to integrate these into external platforms like Tensorflow so they are accessible to developers in the wider AI ecosystem. Since our last report we have released eleven new open-source tools for capabilities including feature visualization for text-based models[6], understanding which parts of an image are most important for ML model predictions[7], and scoring classification models based on how susceptible they are to attacks.[8]

We have also made progress in developing and scaling Model Cards, first introduced as a concept in a research paper in early 2019.[9] Model Cards are like nutrition labels for AI—they explain how a model works and provide information on its intended purpose, performance and known limitations. Over the last year Google has published prototype model cards for the Face Detection and Object Detection features in our Cloud Vision API, and released the Model Cards Toolkit to help internal and external model developers build their own model cards more quickly and easily.

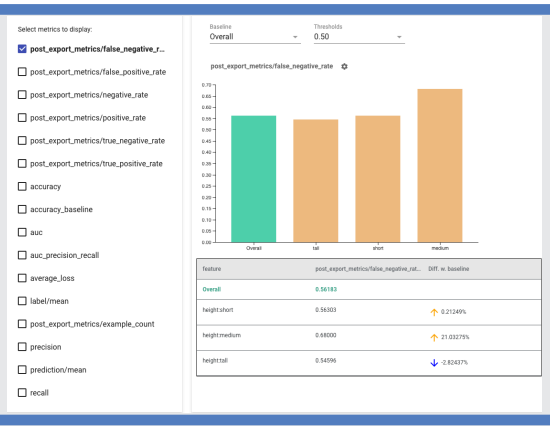
Externally, Model Cards have been adopted by organizations including the Partnership on AI[10] OpenAI[11], and companies like Salesforce.[12] Model cards were named as a best practice in the report *Decision Points in AI Governance* by the University of California, Berkeley's Center for Long-Term Cybersecurity[13] and were also recognized with the 2020 Harvard Kennedy School's Tech and Public Purpose Spotlight Award.

The table highlights some of the most popular tools we have introduced in the past year, with a comprehensive list in the Appendix. Further information on Google's work in this area can be found online at ai.google/responsibilities.

Examples of Responsible AI Tools

Fairness Indicators

In late 2019 we launched a set of fairness testing tools to help developers check their model performance against defined fairness metrics in order to spot potential biases (e.g., a need to reduce the false positive rate for a specific group). Fairness Indicators are currently in use in two dozen internal product pipelines within Google, including YouTube, Jigsaw and Cloud to detect and mitigate potential bias in classifiers. The tool is available to third party developers through Responsible AI with TensorFlow.

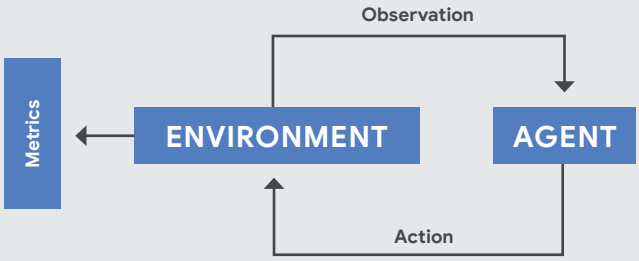


Responsible AI with TensorFlow

Geared toward AI developers who use TensorFlow, this 5-step workflow includes tools for ML fairness, interpretability, privacy, and security. This toolkit complements existing developer processes, so that Responsible AI efforts are directly embedded into a familiar structure.

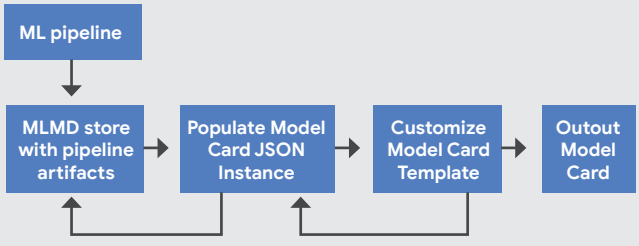
ML Fairness Gym

This open-source set of components for building machine learning model simulations explores the potential long-run impacts of machine learning-based decision systems in social environments.



Model Card Toolkit

To streamline the creation of Model Cards, we released a collection of tools and a Colab tutorial. The toolkit lays out the fields to include in the Model Card and a template for displaying the information. By leveraging Machine Learning MetaData (MLMD), users can pre-populate an instance with many of the fields already filled out.



Internal processes

This year we have continued to develop and improve our review processes to guide research, product, and sales decisions.

Individual teams consider issues as they conduct research and build products, and are complemented by a number of teams that provide tools, infrastructure, and specialized expertise in areas including ethics, fairness, explainability, privacy, security, legal compliance, trust and safety, policy enforcement, risks of abuse of our platforms, and more.

For projects in sensitive fields, or projects for which ethical issues or appropriate mitigations are less clear-cut, project teams can work with a dedicated team to conduct a formal AI Principles review. The diagram provides a high level overview of this process. Other channels, such as office hours, are also available for more consultative discussions on specific AI Principles issues.

The AI Principles Review Process

Any Googler can request a review. Central operations team applies applies relevant AI Principles as ethical frameworks. Internal experts offer specific guidance.

Reviewers systematically consider the AI Principles, and look for examples to apply from previous case studies.

If needed, reviewers consult with experts on mitigation strategy. Product team adjusts approach based on recommended guidance.

The central operations team assesses how to proceed. Or, if challenging issues arise that can affect multiple products, a senior council of Google executives decides how to move forward. The final decision can become a case study.

Any Google employee can request an AI Principles review for a new product, research paper, partnership, or other idea. We also proactively check the pipeline of projects, including research and customer deals, and identify projects with heightened risk that should require a formal review.

The central review team serves as a hub for project teams seeking expert support in carrying out a formal AI Principles review. Team members include user researchers, social scientists, technical experts, ethicists, human rights specialists, policy and privacy advisors, and legal experts on both a full- and part-time basis, which allows for inclusion of diverse perspectives and disciplines. The team also consults with members of Google's Employee Resource Groups (ERGs) to include additional points of view from members of communities that are currently underrepresented in the tech industry.

Some product areas have set up additional review processes to address the product-specific issues they face. For example, Google Cloud has developed a governance

process to address questions that arise in third party enterprise AI deals that differ from consumer-facing Google products. This has two main components: (1) A decision-making body with a clear quorum structure that reviews custom deals and sales engagements for AI Principles risks, and (2) A committee that meets regularly to review AI-related Cloud projects in development, and works with Cloud engineering and product teams to undertake rigorous ethical risk analysis. Reviewers from the central review team also participate to ensure consistency.

From a process perspective, review teams meet at a regular cadence. Reviews are collaborative, with experts working with project teams in an iterative fashion to find solutions to any issues identified. Agreed upon mitigations are signed off on by leadership and become part of product development plans. In the event that there is a challenging decision that requires escalation, a senior council of Google executives conducts a final review and decides how to proceed, including options to significantly change or stop the project.

The central operations team plays a vital role in calibrating decisions across the company, and consolidating a library of case studies as a reference for future decision-making, from which patterns can be identified and precedents established, and which are revisited regularly as new solutions and learnings emerge. The case assessment framework takes into account established internal policies such as Google's Codes of Conduct (for Suppliers and Employees) and relevant environmental, social, and governance (ESG) frameworks, privacy protection policies, and human rights frameworks. External guiding sources include the UN's Universal Declaration of Human Rights, Freedom House's Freedom in the World Report and Transparency International's Corruption Perceptions Index.

External experts play an important role in strengthening our review process as well. For example, we enlisted the nonprofit organization BSR (Business for Social Responsibility) to conduct a formal human rights assessment of the new Celebrity Recognition tool, offered within Google Cloud Vision and Video Intelligence products.[14]

While every review is unique and each case is assessed on its own merits, repeated cases involving a given technology or application can lead to company-wide policies. For example, several AI Principles reviews of text-to-speech (TTS) projects were escalated to the senior executive council, who then determined a Google-wide approach to TTS technologies, including a decision not to open-source TTS models due to the risk that they could be put to nefarious use. These policies ultimately shaped the deployment of TTS in Google's products, including in Google Cloud's Text-to-Speech service and Google Lens.

Google has reviewed hundreds of AI-related projects since our last update. Below are three examples that illustrate key considerations in our reviews.

Face Match in Nest Hub Max

Nest Hub Max is a 10-inch smart home device that helps users make video calls, play music, watch video and TV, see their calendar and reminders, and more. User research identified that Nest users wanted to be able to personalize their settings and experiences, including on shared family devices, beyond the existing Voice Match voice recognition feature. To provide this capability, the Nest team decided to use the camera in the Hub Max to additionally provide a facial recognition capability.

During development of the feature, called Face Match, a key concern was that while facial recognition could enable helpful features for users, it could also feel intrusive in people's homes. In order to ensure that users had control of this feature and their privacy, while also allowing them to personalize their experiences if they wanted to, we designed Face Match to be off by default. The user must expressly turn it on, after receiving clear notice about how their data will be processed by the feature, and can opt out and delete their face data at any time. The face model is encrypted and stored on the Nest Hub Max, and following the setup process, all the face matching occurs locally on-device. That means that after setup, Face Match does not send video or images to Google.

Another concern with Face Match was unfair bias. Some facial recognition systems perform better for some groups than others, often underperforming for women and people of color. The Face Match model went through multiple rounds of testing to identify and address possible sources of gender and skin tone bias.

As a result of these safeguards, multiple people in a home can use Face Match to get personalized help on a shared home device, while ensuring that users are in control of this sensitive feature and how it is deployed in their homes.

Celebrity Facial Recognition

Over the past several years, Cloud AI has been approached by customers in a range of different sectors that want to use Google facial recognition technology. Given the sensitive nature of this technology and the potential for harm, we had never made this technology generally available, but after careful consideration, in 2018 we decided to offer facial recognition products in certain well-defined situations where we felt we could provide valuable services to customers while mitigating risks. As a first implementation of a narrow product offering, we developed a Celebrity Recognition API. Professional media companies had expressed interest in using celebrity facial recognition to understand their huge libraries of content, organize them in new ways, and make them searchable and indexable. While we recognized that this feature would be useful, we also recognized the potential for misuse. During development, the product went through multiple internal reviews and iterations to ensure alignment with the AI Principles, Google's approach to Facial Recognition[15], and our human rights due diligence practices.

As part of the review process, we conducted extensive fairness testing and commissioned external expert guidance from human rights non-profit BSR (Business for Social Responsibility) to conduct a human rights impact assessment (HRIA).[16] The HRIA highlighted the risk that the API could be misused or repurposed for surveillance, and BSR identified a number of recommendations to ensure the API was used as intended which were adopted by the product team. These included establishing specific terms of service for the feature, codifying a narrow definition of "celebrity" and "professionally produced" content, limiting access to vetted customers instead of making the feature publicly available, and providing opt-out and grievance mechanisms. Customers cannot add celebrities to the list (even for private use), and celebrities can also formally opt out if they would like assurances that they will not be recognized by the API.

By integrating these features into the Celebrity Recognition API, we have been able to offer this tool to validated media companies, while limiting the potential for the feature to be misused for surveillance or other purposes.

Deepfake Detection

Deepfakes are a type of photo-realistic synthetic media, in which a person's image or video image is manipulated to generate or replace that of another, which can be used nefariously to create mis- or disinformation about a real person. Online platforms use detectors to identify deepfakes in online images in video, but these detectors are not infallible, and our research teams wanted to understand how malicious actors might try to evade our detectors and disguise manipulated content. Early in 2020, a Google-UC Berkeley research collaboration identified five types of attacks conducted on a state of the art machine-learning classifier used to detect deepfakes. The Google researcher requested an AI Principles review to help guide how this information should be used and published.

A primary concern was that publicly highlighting vulnerabilities and describing specific techniques could help malicious actors to evade detectors. On the other hand, deepfakes threaten to spread misinformation, disinformation, and defamation and erode trust in journalism and media, and sharing this research could improve the sensitivity and robustness of deepfake detectors.

Ultimately, the review team determined that the paper[17] describing the study and outlining key takeaways for teams building deepfake detectors could help advance the field of robust deepfake detection. However, we decided not to open-source the model used by the researchers in order to ensure that it could not be used to evade deepfake detectors. The paper was approved and presented at the CVPR workshop on Media Forensics.

Community outreach and exchange

Google remains committed to engaging with external experts and civil society stakeholders around the world. Establishing norms and best practices for responsible AI development will succeed only as a community endeavour, so it is vital to share learnings and receive feedback on our efforts from the wider community.

This year, we have continued to sponsor programs on inclusion, diversity, and equity in AI research and product development, including WiML (Women in ML) and Latinx in AI; many of these organizations are also co-founded and co-led by Google employees, such as Black in AI and Queer in AI. And we continue to partner with the Algorithms for Opacity Group (AFOG) at UC Berkeley, an interdisciplinary research group that bridges disciplinary boundaries to support the equitable development of AI systems.

Although in 2020 the COVID-19 pandemic has limited our capacity to attend and stage in-person gatherings, we have sought to migrate to virtual formats wherever possible. For instance, we developed an in-person ML for Policy Leaders course, and converted it into a two hour interactive online workshop. Since May 2020, we have run more than a dozen virtual workshops with more than 180 policymakers and 60+ organizations across the US, the EU, and LATAM.

We've also made a concerted effort to offer hands-on learning opportunities for communities who have been historically underrepresented in the technology industry. Our global community educational outreach efforts have included TensorFlow bootcamps offering introductory machine learning skills across India and Latin America and a May 2020 ML Bootcamp for mobile developers, targeting communities who may lack strong broadband infrastructure or access to desktop computing. The ML Bootcamp was live streamed on YouTube to 4,500 participants in 50 nations across Asia, Europe, and Africa, and offered remote mentorship with developer advocates from Google. Additionally, in January 2020, we launched Bangkit, an intensive, free ML training program to help Indonesian developers build technical ML skills, as well as more general "soft skills" that can help them advance their career in the technology sector.

The following tables provide an extended summary of global engagement highlights from the last year.

Community workshops

APAC Machine Learning Study Jams: Hands-on ML training in Hong Kong,[18] Indonesia[19], Japan,[20] and Korea,[21] among other countries across the region.

AI Bootcamp in Taipei in partnership with the Taiwan Ministry of Science and Technology: Event held to further our commitment to train 30K Taiwanese students, researchers and developers through open fundamental ML courses, including the Google Developers Machine Learning Crash Course.[22]

AI Everything Summit sponsored by the UAE government and hosted a three-day workshop, featuring recommended best practices on Responsible AI.

AI Professional development in Australia: Training focused on introducing Artificial Intelligence and Machine Learning in the classroom and was hosted by our partner, the University of Adelaide CSER group. In June 2020, Google Australia hosted the first virtual AI professional development (PD) workshop for 80 teachers on Grow With Google OnAir.[23]

Black in AI: Participation and funding for workshops and programs supporting Black individuals working in AI.[24]

LatinX in AI: Participation and funding for workshops and programs supporting Latinx individuals working in AI.[25]

WiML Workshops: Participation and support for the Women in Machine Learning (WiML) workshops, which facilitate exchanges between female faculty, research scientists, and graduate students in the machine learning community.[26]

Queer in AI workshops: Participation and funding for workshops and social events raising awareness of queer issues in AI and fostering a community of queer researchers.[27]

CtrlZ.AI: An event on mitigating the harms of data-driven algorithmic systems in conversation with artists, activists, and diverse communities affected by algorithmic systems.[28]

FAccT tutorials and workshops: Instructive guidance-based sessions at FAccT (ACM Conference on Fairness, Accountability, and Transparency), each attended by ~100 people.[29]

Global People + AI Guidebook UX workshops: More than 170,000 people from 195 countries have read the People + AI Guidebook[30], which is designed to help UX designers and product managers adopt a human-centered approach to AI development. It has been incorporated into responsible design curriculum at major universities, including Stanford and Oxford, and been the basis of numerous workshops including at the Gates Foundation and Public Health England, Interaction 20, a World Economic Forum Global Shapers event, DevFest in Brazil, and a WiML event in Ghana.

Grace Hopper Celebration of Women in Computing 2019 cross-industry panel, "Embedding Social Responsibility into the Engineering Lifecycle".

Kids AI Programming Competition: Exposed students in Japan to CS/AI education at home. The contest targets primary and middle school students who want to express their creativity through Scratch programming and AI models developed on TeachableMachine.[31]

MENA AI Hackathon: The Google Dev Ecosystem, in partnership with InstaDeep, a Tunisia-founded enterprise AI startup, organized the largest AI Hackathon in the history of the MENA region with 1,000 attendees from over 15 countries, with 37% women attendees and 55% women speakers were women.

MENA Study Jams Training program: Since 2019 we have gained 40,000 TensorFlow learners and Assistant learners. Committed to train +100,000 Developers yearly on ML, Flutter, Android and Google cloud technologies.

ML Fairness Workshops at CVPR (Computer Vision and Pattern Recognition) conference.

ML Practicum: Fairness in Perspective API: This is a free online course[32] that illustrates how the Jigsaw team addressed online harassment in partnership with Google's Counter-Abuse Technology team by developing the Perspective API. The practicum also offers hands-on use of Fairness Indicators to evaluate ML models and to help mitigate unintended bias in training data.

Online ML Bootcamp for mobile developers: This event was led by Singapore-based Googlers and live-streamed on YouTube to 4,500 participants in 50 nations across Asia, Europe, and Africa.

Practical ML for Developing Countries Workshop at ICLR (International Conference on Learning Recognition.[33]

Responsible AI Cloud customer workshops in the United States and United Kingdom: Shared Google's open-source Responsible AI tools with 7 clients in healthcare, banking and education, resulting in technology pilots. When complete, this training will be offered through the Partner Advantage Program at no cost.

Responsible AI Panel at Google's AI for Social Good Impact Challenge Summit: Discussion on ML fairness with external panelists from Crisis Text Line, TalkingPoints, Fondation Médecins Sans Frontières.

Socially Responsible NLP tutorial at the Web Conference 2019: Provided an overview of real-world NLP technologies and their potential ethical implications.[34]

Societal Context Research presentations at AFOG and Data & Society: With U.C. Berkeley's Algorithmic Fairness and Opacity Working Group (AFOG) and Data & Society.

Solve with AI: A half day conference on AI for social good in Tokyo, featuring Googlers and several AI for Social Impact challenge grantees.

Solving the Algorithm: Women in Machine Learning in Accra, Ghana in December 2019, which offered free machine learning training for 100 African women.

TensorFlow roadshows: ML bootcamps introductory machine learning skills at Google offices (pre-COVID 19 pandemic) across India and Latin America.

What-If Tool Workshop: Introduced leading researchers and industry technologists to use What-If to identify dataset imbalances and analyze human-centered ML models from a fairness perspective.[35]

Bangkit Indonesia: Intensive educational program designed with leading Indonesian companies and universities to train 300 Indonesian participants in ML concepts and tools.

Competitions and grant programs

AI Social Impact Challenge: We selected 20 grantees applying AI to address social and environmental challenges out of a pool of 2,602 applications from 119 countries. Grantees receive a combined $25M in grants from Google.org, along with additional credit and consulting from Google Cloud, mentoring from Google AI experts, and training at a customized program from Google Developers Launchpad Accelerator.[36]

Cassava vision Kaggle challenge: We launched a competition encouraging the use of computer vision to identify cassava leaf disease, helping to protect a key food security crop for small-holder farmers in Africa. The goal is to train a model to classify a given image into 4 disease categories or a healthy leaf, using a dataset of 9,436 annotated images and 12,595 unlabeled images of cassava leaves.[37]

Google.org Impact Challenge on Safety: Google.org launched a €10 million grant fund for European safety researchers. Each grantee receives funding of €50 thousand to €1 million and support from Google researchers.[38]

Award for Inclusion Research Program: A global program to recognize and support academic research in computing (or technology) that addresses the needs of underrepresented populations. We selected 16 grantees working on topics including diversity and inclusion, algorithmic bias, education innovation, health tools, accessibility, gender bias, AI for social good, security, and social justice.[39]

Other engagements

OECD Network of Experts on AI (ONE AI): We are formal members of this network and active participants in working groups to help in drafting guidance on the implementation of the OECD's AI Principles, sharing learnings from Google's experience.

National AI Research Institute: We have partnered with the National Science Foundation in the United States to establish a National AI Research Institute to advance studies of Human-AI Interaction.

Partnership on AI: We continue to serve on the board and as a member of the Partnership on AI, a multi-stakeholder organization that studies and formulates best practices on AI technologies. As an example of our work together, the Partnership on AI is developing best practices that draw from our Model Cards proposal as a framework for accountability among its member organizations.

UNGA panel: We co-hosted a virtual panel at the United Nations General Assembly on how AI can be used to accelerate progress towards the UN Sustainable Development Goals, which featured representatives from three Google.org grantees (Fondation MSF, Lacuna Fund, and Chequeado).[40]

AI standards: Google is a founding member of the International Organization for Standardization's (ISO) subcommittee 42 on artificial intelligence, including contributing to draft working documents on AI ethics, risk management and bias.

Equitable AI Research Roundtable (EARR): EARR is a participatory model we developed for including external community-based researchers in a collaborative ML Fairness research process.

Journalism AI: A year-long collaboration between Polis, the international journalism think tank at the London School of Economics and Political Science, and the Google News Initiative, on the potential for AI to influence the future of journalism.[41]

ML for Policy Leaders: A workshop on machine learning targeted towards policymakers, ML for Policy Leaders. The focus of the workshop is primarily on ethical challenges and governance, with detailed case studies based on real-world applications. It also includes interactive simulations to assist in exploring issues related to Explainability, Fairness, Privacy, and Security.

Submissions to government consultations: During 2020 we have responded to numerous calls for input from governments around the world seeking to explore AI governance issues, including in the US, the EU and India.

Responsible AI presentation at TensorFlow Dev Summit (live streamed): Presented "Responsible AI with TensorFlow: Fairness and Privacy," which featured a fairness-aware ML workflow, the Fairness Indicators tool, and infrastructure pieces to help train a model in a privacy preserving manner.[42]

ML Perf: A collaboration between companies and academic researchers to build fair and useful benchmarks for measuring training and inference performance of ML hardware, software, and services.[43]

Conclusion

We recognize the need for a principled approach to developing and applying AI. As our CEO Sundar Pichai said in January, it's crucial that companies like ours not only build promising new technologies, but also harness them for good—and make them available for everyone.[44] As we noted last year, no single company, community, or country can address the challenges and harness the opportunities that AI presents. While we have made continued progress in 2020, there remains significant work to do within the company as well as across the ecosystem as a whole. We are committed to sharing our learnings and tools to support responsible AI innovation, including through regular updates to the Responsibilities, Tools and Education sections of the Google AI website (ai.google).

Appendix: Research publications and tools

Research publications

- **A Partial Break of the Honeypots Defense to Catch Adversarial Attack.** Nicholas Carlini. September 2020.
- **Bringing the People Back In: Contesting Benchmark Machine Learning Datasets.** Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, Morgan Klaus Sheuerman. ICML 2020.
- **Building Healthy Recommendation Sequences for Everyone: A Safe Reinforcement Learning Approach.** Ashudeep Singh, Yoni Halpern, Nithum Thain, Konstantina Christakopoulou, Ed H. Chi, Jilin Chen, and Alex Beutel. ACM FAccT 2020.
- **CAT-Gen: Improving Robustness in NLP Models via Controlled Adversarial Text Generation.** Tianlu Wang, Xuezhi Wang, Yao Qin, Ben Packer, Kang Li, Jilin Chen, Alex Beutel, Ed Chi. EMNLP 2020.
- **Characterizing Private Clipped Gradient Descent on Convex Generalized Linear Problems.** Shuang Song, Om Thakkar, Abhradeep Thakurta. June 2020.
- **Closing the Accountability Gap: Defining a Framework for Internal Algorithmic Auditing.** Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, Parker Barnes. ACM FAT* (Fairness, Accountability and Transparency) conference 2020.
- **Debugging Tests for Model Explanations.** Julius Adebayo, Michael Muelly, Ilaria Liccardi, Been Kim. ICML 2020.
- **Deontological Ethics By Monotonicity Shape Constraints**. Serena Wang, Maya Gupta. AISTATS 2020.
- **Diversity and Inclusion Metrics for Subset Selection.** Margaret Mitchell, Nyalleng Mooros, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, Jamie Morgenstern. AAAI/ACM Conference on AI, Ethics, and Society (AIES), ACM, 2020.
- **Enabling certification of verification-agnostic networks via memory-efficient semidefinite programming.** Sumanth Dathathri, Krishnamurthy Dvijotham, Alexey Kurakin, Aditi Raghunathan, Jonathan Uesato, Rudy Bunel, Shreya Shankar, Jacob Steinhardt, Ian Goodfellow, Percy Liang, Pushmeet Kohli. October 2020.
- **Encode, Shuffle, Analyze Privacy Revisited: Formalizations and Empirical Evaluation.** Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Shuang Song, Kunal Talwar, Abhradeep Thakurta. January 2020.
- **Estimating Training Data Influence by Tracking Gradient Descent.** Garima Pruthi, Frederick Liu, Mukund Sundararajan, Satyen Kale. February 2020.
- **Evaluating Attribution for Graph Neural Networks.** Benjamin Sanchez-Lengeling, Jennifer N. Wei, Brian K. Lee, Emily Reif, Peter Y. Wang, Wesley Wei Qian, Kevin McCloskey, Lucy Colwell, Alexander B. Wiltschko. NeurIPS 2020.
- **Explainability Case Studies.** Ben Zevenbergen, Allison Woodruff, Patrick Gage Kelley. CSCW Workshop on Ethics in Design 2020.
- **Advances and Open Problems in Federated Learning.** Peter Kairouz, H. Brendan McMahan, Brendan Avent, Afurelien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, et al. December 2019.

- **Extracting Neural Networks through Cryptanalytic Techniques.** N. Carlini, M. Jagielski, I. Mironov. Crypto 2020.
- **Fairness without Demographics through Adversarially Reweighted Learning.** Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, Ed H. Chi. NeurIPS 2020.
- **Fairness with Overlapping Groups.** Forest Yang, Moustapha Cisse, Sanmi Koyejo. NeurIPS 2020.
- **Federated Heavy Hitters Discovery with Differential Privacy.** Wennan Zhu, Peter Kairouz, Brendan McMahan, Haicheng Sun, Wei Li. AISTATS 2020.
- **Fundamental Tradeoffs between Invariance and Sensitivity to Adversarial Perturbations.** Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, Jörn-Henrik Jacobsen. ICML 2020.
- **Generative Models for Effective ML on Private, Decentralized Datasets.** With Federated Assistant. Sean Augenstein, Brendan McMahan, Daniel Ramage, Swaroop Ramaswamy, Peter Kairouz, Mingqing Chen, Rajiv Mathews, Blaise Aguera y Arcas. ICLR 2020.
- **High Accuracy and High Fidelity Extraction of Neural Networks.** Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, Nicolas Papernot. Usenix 2020.
- **Human Evaluation of Models Built for Interpretability.** Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J. Gershman, Finale Doshi-Velez. AAAI Conference on Human Computation and Crowdsourcing 2019.
- **Label-only membership inference attacks.** Christopher A. Choquette Choo, Florian Tramer, Nicholas Carlini, Nicolas Papernot. July 2020.
- **Learning to Diversify from Human Judgements.** Emily Denton, Hansa Srinivasan, Dylan Baker, Jilin Chen, Alex Beutel, Tulsee Doshi, Ed H. Chi. CHI2020.
- **Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning.** E. Jo; T. Gebru. (ACM FAccT). January 2020.
- **Measuring Robustness to Natural Distribution Shifts in Image Classification.** Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, Ludwig Schmidt. NeurIPS 2020.
- **Obliviousness Makes Poisoning Adversaries Weaker.** Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Abhradeep Thakurta. March 2020.
- **On Adaptive Attacks to Adversarial Example Defenses.** Florian Tramer, Nicholas Carlini, Wieland Brendel, Aleksander Madry. NeurIPS 2020.
- **On Completeness-Aware Concept-Based Explanations in Deep Neural Networks.** Chih-Kuan Yeh, Been Kim, Sercan O. Arik, Chun-Liang Li, Tomas Pfister, Pradeep Ravikumar. NeurIPS 2020.
- **Pairwise Fairness for Ranking and Regression.** Harikrishna Narasimhan, Andy Cotter, Maya Gupta, Serena Lutong Wang. 33rd AAAI Conference on Artificial Intelligence 2020.
- **Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics Approach.** Donald Martin Jr., Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, William S. Isaac. May 2020.
- **Practical Compositional Fairness: Understanding Fairness in Multi-Component Ranking Systems.** Xuezhi Wang, Nithum Thain, Anu Sinha, Ed H. Chi, Jilin Chen, Alex Beutel. November 2019.
- **Privacy Amplification via Random Check-Ins.** Borja Balle, Peter Kairouz, H. Brendan McMahan, Om Thakkar, Abhradeep Thakurta. TPDP 2020.
- **ReMixMatch: Semi-Supervised Learning with Distribution Matching and Augmentation Anchoring.** David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, Colin Raffel. ICRL 2020.

- **Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing.** Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, Emily Denton. AAAI/ACM AI Ethics and Society conference 2020.
- **Social Biases in NLP Models as Barriers for Persons with Disabilities.** Ben Hutchinson, Emily Denton, Kellie Webster, Stephen Craig Denuyl, Vinodkumar, Prabhakaran, Yu Zhong. ACL 2020.
- **Stateful Detection of Black Box Adversarial Attacks.** David Wagner, Nicholas Carlini, Steven Chen. 1st Workshop of Security and Privacy in Artificial Intelligence 2020.
- **Tempered Sigmoid Activations for Deep Learning with Differential Privacy.** Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, Úlfar Erlingsson. July 2020.
- **The Flajolet-Martin Sketch Itself Preserves Differential Privacy: Private Counting with Minimal Space.** Adam Smith, Shuang Song, Abhradeep Thakurta. NeurIPS 2020.
- **The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models.** Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, Ann Yuan. EMNLP 2020.
- **Thieves on Sesame Street! Model Extraction of BERT-based APIs.** Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, Mohit Iyyer. ICLR 2020.
- **Toward a better trade-off between performance and fairness with kernel-based distribution matching.** Flavien Prost, Hai Qian, Qiuwen Chen, Ed H. Chi, Jilin Chen, Alex Beutel. October 2019.
- **Towards a Critical Race Methodology in Algorithmic Fairness.** Alex Hanna, Emily Denton, Andrew Smart, Jamila Smith-Loud. ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) 2020.
- **Why Reliabilism Is not Enough: Epistemic and Moral Justification in Machine Learning.** A Smart, L James, B Hutchinson, S Wu and S Vallor. AAAI AIES. Feb 2020.
- **Evading Deepfake-Image Detectors with White- and Black-Box Attacks.** Nicholas Carlini, Hany Farid. April 2020.
- **Robustness to large-step adversarial manipulations for a subset of features.** Aradhana Sinha, Amer Sinha, Ben Packer, Xuezhi Wang, Nithum Thain, Corey Lane, Ed Chi, Alex Beutel, Jilin Chen. KDD 2020.
- **Dimension Independence in Unconstrained Private ERM via Adaptive Preconditioning.** Peter Kairouz, Mónica Ribero, Keith Rush, Abhradeep Thakurta. August 2020.

Tools

- **Attack Library Privacy Attack Test.** New mode in TensorFlow Privacy to help developers assess privacy properties of classification models. The new test produces a score, with a high score indicating susceptibility to attack. https://github.com/tensorflow/privacy
- **COVID-19 Research Explorer.** Semantic search UI for the COVID-19 Open Research Dataset of over 45,000 journal articles and preprints. This tool can help scientists and researchers efficiently pore through articles for answers to COVID-19-related questions. https://covid19-research-explorer.appspot.com/
- **Explainable AI with Google Cloud.** Tools and frameworks to help develop interpretable and inclusive machine learning models and deploy them with confidence, including guidance on known limitations. https://cloud.google.com/explainable-ai/
- **Fairness Indicators.** This tool helps check model performance against defined fairness metrics, e.g., reducing the false positive rate for a specific group or groups. https://github.com/tensorflow/fairness-indicators

- **Feature Visualization for Text.** This tool provides visual insight into how we adapted the techniques of feature visualization to text-based models. https://pair-code.github.io/interpretability/text-dream/explainable/
- **Language Interpretability Tool (LIT):** A visual, interactive, model-understanding tool for NLP models. LIT helps developers determine what datapoints a language model performs poorly on, and why a model made a certain prediction. https://pair-code.github.io/lit/
- **MinDiff for ML Fairness:** The first new tool in our new Model Remediation Library, MinDiff helps model developers to balance error rates for different users or contexts of use, helping to mitigate potentially unfair outcomes for vulnerable populations. https://www.tensorflow.org/responsible_ai/model_remediation/
- **ML Fairness Gym:** A set of components for building simple simulations that explore the potential long-run impacts of deploying machine learning-based decision systems in social environments. https://github.com/google/ml-fairness-gym
- **Model Cards Toolkit.** A collection of tools and Colab tutorial to help model developers build Model Cards efficiently by leveraging Machine Learning MetaData (MLMD) to prepopulate fields in their Model Card. https://github.com/tensorflow/model-card-toolkit
- **Responsible AI with TensorFlow.** A consolidated toolkit for third party developers on TensorFlow to build ML fairness, interpretability, privacy, and security into their models. https://www.tensorflow.org/resources/responsible-ai
- **XRAI.** Method that helps developers better understand which regions of an image were most important for the predictions the model made. Recently integrated into Cloud's Explainable AI platform. http://bit.ly/XRAI_Tool

Endnotes

1 https://github.com/tensorflow/privacy

2 https://arxiv.org/pdf/1910.11779.pdf

3 https://arxiv.org/pdf/2006.13114.pdf

4 https://arxiv.org/pdf/2006.13114.pdf

5 https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

6 https://pair-code.github.io/interpretability/text-dream/explainable/

7 http://bit.ly/XRAI_Tool

8 https://github.com/tensorflow/privacy

9 https://arxiv.org/abs/1810.03993

10 https://www.partnershiponai.org/the-partnership-on-ai-launches-multistakeholder-initiative-to-enhance-machine-learning-transparency/

11 https://github.com/openai/gpt-3/blob/master/model-card.md

12 https://help.salesforce.com/articleView?id=mc_anb_einstein_messaging_insights_model_card.htm&type=5

13 https://cltc.berkeley.edu/2020/05/05/decision-points-in-ai-governance/

14 https://www.bsr.org/en/our-insights/blog-view/google-human-rights-impact-assessment-celebrity-recognition

15 https://ai.google/responsibilities/facial-recognition/

16 https://www.bsr.org/en/our-insights/blog-view/google-human-rights-impact-assessment-celebrity-recognition

17 https://arxiv.org/abs/2004.00622

18 https://events.withgoogle.com/ml-study-jam-basic-hk/

19 https://developers.google.com/machine-learning/crash-course?hl=id

20 https://developers-jp.googleblog.com/2020/04/ml-study-jams-vol4-machine-learning.html

21 https://byeongsupark.github.io/en/blog/google-ml-studyjam/machine-learning-studyjam

22 https://developers.google.com/machine-learning/crash-course

23 https://growonairau.withgoogle.com/events/education

24 https://blackinai2020.vercel.app/

25 https://www.latinxinai.org/

26 https://wimlworkshop.org/

27 https://sites.google.com/view/queer-in-ai/home?authuser=0

28 http://www.ctrlz.ai/

29 https://fatconference.org/index.html

30 https://pair.withgoogle.com/guidebook/

31 https://g.co/kidsaicontestjp

32 https://developers.google.com/machine-learning/practica/fairness-indicators

33 https://pml4dc.github.io/iclr2020/

34 https://www2019.thewebconf.org/workshops-and-tutorials

35 https://pair-code.github.io/what-if-tool/fat2020.html

36 http://services.google.com/fh/files/misc/accelerating_social_good_with_artificial_intelligence_google_ai_impact_challenge.pdf

37 https://www.kaggle.com/c/cassava-disease/data

38 https://impactchallenge.withgoogle.com/safety2019/process

39 https://research.google/outreach/air-program/

40 https://www.youtube.com/watch?v=Q0SPnVN_XBI&feature=youtu.be

41 https://www.blog.google/outreach-initiatives/google-news-initiative/how-ai-could-shape-future-journalism/

42 https://www.youtube.com/watch?v=UEECKh6PLhI

43 https://mlperf.org/

44 https://www.ft.com/content/3467659a-386d-11ea-ac3c-f68c10993b04

Google