Google

AI Principles 1-Year Progress Update

Table of contents

Overview 2

Culture, education, and participation 4

Technical progress 6

Internal processes 8

Community outreach and exchange 11

Conclusion 14

Appendix: Research publications and tools 15

End notes 19

Overview

AI offers incredible potential to empower people, promote innovation, and widely benefit current and future generations. But, like any transformational technology, it also raises important challenges that need to be addressed clearly and thoughtfully. That is why one year ago we published the Google AI Principles (see box) as a charter guiding the responsible development and use of artificial intelligence in Google's business. In the past year, we have focused on building the processes, teams, tools and training necessary to operationalize the Principles. This report provides an update on our progress.

Of course, many experts around the world are working on issues of AI and the impact it will have on society. Ultimately it is up to countries and communities to choose how they want to harness the benefits of AI, and to establish the right frameworks for its development. Google is committed to playing our part by contributing ideas, sharing best practice learnings, and being transparent about our progress. The aim of this report is to provide a pointer to relevant technical work (for practitioners) and give a sense (for a general audience) of the breadth of our activity and how we are translating our principles into practice.

First, we are working to incorporate the Principles in the day-to-day working processes of all teams engaged in AI-related development. We have put significant effort into culture, education, and participation. For example, we currently host a variety of speaker series on responsible AI, and are piloting a custom interactive workshop on tech ethics and our AI Principles. Additionally, we've developed a technical module on machine learning fairness that is included as part of our Machine Learning Crash Course, used to train thousands of people internally and externally.

Second, technical progress plays a fundamental role in identifying tools and solutions for responsible AI. This year, for example, we updated Google Translate to reduce gender bias[1]; launched the What-If Tool[2], which enables users to visualize biases and the effects of various fairness constraints; and expanded on federated learning[3], which allows developers to train and deploy AI models without data leaving the device. We also continue to update our Responsible AI Practices[4] quarterly to reflect the latest technical ideas and work at Google.

Third, as previously described[5], we have established a process for reviewing our products and services to promote thoughtful consideration of tradeoffs and balanced decision making. In this report we also share two project review cases to illustrate our decision making process.

Finally, promoting the responsible development and use of AI requires collaboration across sectors and we regularly participate in discussions, gather inputs, and solicit feedback from external experts, policymakers and civil society. In the past year, we have engaged

in more than 100 workshops, roundtables, research conferences, panels, working groups, closed-door dialogues, and summit events on these topics, engaging with thousands of stakeholders in Europe, Asia Pacific, Africa, Latin America, and North America.

As we continue to explore the important questions raised by the development and deployment of advanced technologies, we are committed to sharing our learnings with others and eager to hear their learnings in turn. Our goal is to contribute to an open dialogue that helps the whole community learn and make progress.

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

We encourage all of our employees to participate in the responsible use and development of AI. This includes both understanding AI technologies and tools and gaining familiarity with issues in how they are applied. For example, more than 10,000 Googlers took training on how to address unfair bias in machine learning, and more than 40,000 students have completed the Fairness module that is part of our Machine Learning Crash Course.

In Q4 2018, our internal education program team launched a training course, 'Tech Ethics in Practice,' modeled on materials developed by the Markkula Center for Applied Ethics at Santa Clara University. The course helps participants understand the responsibilities that accompany technical work by exploring fundamental concepts, tools, and practical approaches to applying ethical considerations into the design, development, and implementation of technology. It includes a review of our AI Principles, case studies for analysis and discussion, and pointers to internal resources for questions and counsel. It also offers guidance for fostering constructive cross-functional dialogue and decision making.

While some trainings have targeted specific roles, such as Product Managers or AI specialists, most trainings have included cross-functional groups, typically a mix of employees in Engineering, Research, Policy, Communications, Legal, Sales, People Operations, Trust & Safety and management. Most recently, the team developed a version of the training customized specifically for those in key leadership roles at Google. The focus of that course is modeling ethical practice and developing the right incentives, team dynamics, performance metrics, resources, and formal processes necessary to support and reward responsible innovation within the teams they lead. The leadership training has also been provided to senior members of the AI Principles' governance structure who advise on especially complex, sensitive, or challenging decisions related to our AI Principles.

The training has been offered in our Mountain View, Sunnyvale, Munich, and Zurich offices, with more sessions planned for later this year. Thus far, in-person trainings have been delivered to around 200 Googlers across all core product areas, including more than fifty in senior leadership.

Also this year, the Education team began production of a self-service video version of the Tech Ethics in Practice training that will be offered to all Googlers worldwide, scheduled for release later in 2019. The course is divided into four modules: concepts and history, how ethicists think of technology, lenses and frameworks for decision making, and techniques for applying tech ethics at Google. It will combine lectures with self-guided exercises to promote practical application. Our goal is for thousands of employees to take the self-service training by the end of the year.

In parallel, we've hosted a variety of speaker series on responsible AI, bringing in internal and external experts to speak on topics like fairness, explainability, privacy, and security. Additionally, we've developed a technical module on machine learning fairness that is included as part of our Machine Learning Crash Course, used to train thousands of people internally and externally.

Implementing the AI Principles is a company-wide effort, and throughout the past year, more than forty Google teams have been involved across product, engineering, operations, research, privacy, legal, and sales. Some of these teams include Trust & Safety, Research, Cloud AI, Privacy and Security, UX, Public Policy, Education, Product Inclusion, and Human Rights & Social Impact.

Technical progress

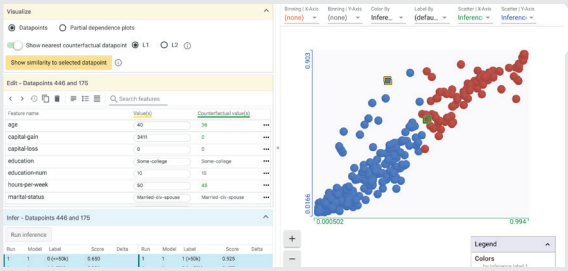
The development of AI is creating new opportunities to improve the lives of people around the world, from business to healthcare to education. It is also raising new questions about the best way to build interpretability, privacy, security and fairness into these systems. These questions are active areas of research and development, and we are committed to sharing knowledge, research, tools, datasets, and other resources with the larger community.

During the past year we have published a total of 79 research papers on AI ethics, fairness, explainability, privacy, and security; and developed 12 new open-source tools (see Appendix). Some of this work is highlighted in the tables below, and a full list of Google's work in this area can be found at: https://ai.google/responsibilities/responsible-ai-practices/

Fairness tools and research

Open Images eXtended[6]: A new branch of Google's Open Images dataset, to expand the diversity of cultures and people represented in imagery training data. Includes a prototype data card[7], providing transparency on the background, characteristics, and composition of the dataset.

Model Cards for Model Reporting[8]: A modular, extensible framework for transparent reporting of ML models' ethical considerations and fairness evaluations.



What-If Tool[9]: An interactive tool built into TensorBoard, and usable in Jupyter and Colaboratory[10] notebooks, to analyze model performance on a dataset and observe the effects of changes to input data or model constraints; ena bles users to visualize biases and the effects of various fairness constraints as well as compare performance across multiple models.

Ensuring fairness in ML to advance health equity[11]: This paper provides an overview of the interplay between ML fairness and health disparities.

Measuring unintended bias in text classification[12]: Research, with accompanying tutorial[13] and Kaggle competition[14], that demonstrates a suite of threshold-agnostic metrics to find new and potentially subtle unintended bias in existing models.

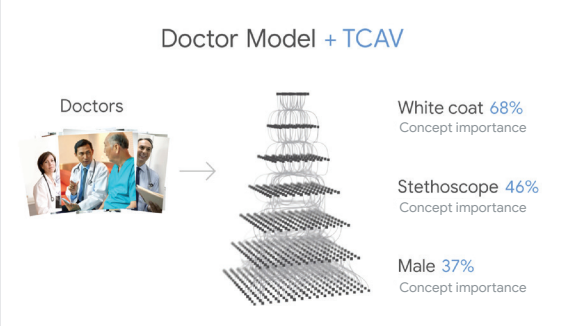
Fairness in Recommendation Ranking through Pairwise Comparisons[15]: Offers multiple fairness metrics to evaluate and understand ranking and recommendation systems, as well as approaches to improve those metrics during model training.

Counterfactual fairness in text classification through robustness[16]: Outlines and compares multiple approaches for addressing counterfactual fairness issues in text models.

History of Fairness[17]: Provides details on parallels between current algorithmic fairness work and test fairness work in the 1960s and 1970s, with directions forward learned from the past.

Machine Learning Crash Course fairness module[18]: Developed an Intro to Fairness module to our machine learning crash course, available for anyone to take for free.

Interpretability tools and research



TCAV[19]: Continued progress in quantifying the importance of a user-chosen high-level concept (e.g., gender) for model classification decisions; and application into a tool to help doctors[20] navigate medical images. TCAV was also awarded a UNESCO Netexplo award[21].

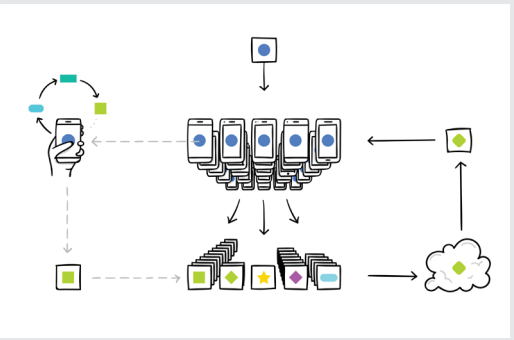
Interpreting Black Box Predictions using Fisher Kernels[22]: Identifies training examples which are most responsible for a given set of predictions.

To Trust or Not to Trust a Classifier[23]: Research and open-source library[24] documentation method to determine if the prediction from a classifier can be trusted for more responsible use of machine learning models.

Human-in-the-Loop Interpretability Prior[25]: A framework to learn more interpretable models for a given end task by asking humans which models are more interpretable during the model training.

Privacy & Security tools and research

Federated learning[26]: A distributed machine learning approach in which a fleet of devices coordinates to train a shared global model from locally-stored data. Originally described in this research paper[27], the technique is now used to power mobile keyboard prediction[28] in GBoard and open-sourced as TensorFlow Federated[29].



TensorFlow Privacy[30]: An open source library for developers to train ML models with differential privacy, and for researchers to advance the state of the art in ML with strong privacy guarantees.

On Evaluating Adversarial Robustness[31]: Overview of methodological foundations, and practices and methods for evaluating defenses to adversarial examples.

General tools

People + AI Guidebook[32]: A toolkit of methods and decision-making frameworks for teams to build human-centered AI products.

Creatability[33]: A set of experiments made in collaboration with creators and allies in the accessibility community, exploring how creative tools can be made more accessible using web and AI technology.

Dataset Search[34]: A new search tool for anyone to find useful and interesting datasets on the Web, from various scientific disciplines, such as environmental or social science, to government or economic data from around the world.

Internal processes

We make our product and business decisions around AI through a series of assessments that ensure rigor and consistency in our approach across product areas and geographies.

Our development teams consider issues as they build products, and are complemented by a number of teams that provide specialized expertise in areas including privacy, legal compliance, trust and safety, policy enforcement, risks of abuse of our platforms, and more. Our responsible innovation team handles day-to-day operations and assessments of projects with particular relevance to AI. The team includes 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 is responsible for developing standards, establishing precedents, and consolidating resources to be used in decision-making.

We also draw on a group of senior experts from a range of disciplines across Alphabet who provide technological, functional, and application expertise. This group provides input on a case-by-case basis, allowing flexibility and responsiveness to rapid advances in technological capabilities and unique applications. These experts have organization-level visibility, accountability and authority to make product or technology launch decisions. Senior Google executives review complex and challenging issues, helping set policies that affect multiple products and technologies.

While our assessments vary by context—AI work at Google ranges from publishing research to launching commercial APIs—we assess each project against the full set of Principles, taking into account the likelihood of potentially beneficial and/or negative outcomes. Learning from our own experiences and those shared by others, we continue to iterate and improve on our implementation processes.

Google has reviewed hundreds of AI-related projects in the past year. Below are three examples to illustrate the types of considerations and trade-offs we have explored.

Sample-Efficient Adaptive Text-to-Speech Research

A research group within Google wrote an academic paper detailing an efficient text-to-speech (TTS) network. This system addresses a major challenge in AI research: systems often need to be retrained from scratch, sometimes with huge amounts of data, to take on even slightly different tasks. In the paper, the researchers demonstrate a more efficient approach, allowing a system to be trained once with a high volume of training data, and then adapted to different contexts with much less time and data.

While this project has accessibility benefits, such as helping individuals with voice disabilities, ALS, or tracheotomies communicate following surgery, we recognized the potential for harmful applications of text-to-speech technologies, such as synthesis of particular voices for deceptive purposes or violation of voice-activated security protocols. We reviewed this project to assess whether publishing the paper might result in overall

harm, or unfair bias for some communities (e.g., whether voices of certain ages, accents, or genders are harder or easier to reproduce with this model).

Ultimately we determined that the technology described in the paper had limited potential for misuse for several reasons, including the quality of general and unique-individual training data required to make it work. Arbitrary recordings from the internet would not satisfy these requirements. In addition, there are enough differences between samples generated by the network and speakers' voices for listeners to identify what's real and what's not.

We accordingly concluded that this paper was overall in line with our AI Principles. We plan to continue to publish recent research results on TTS advances and to include protected and privacy-preserving models on personal phones, but we do not plan to open source the code to create text-to-speech models from arbitrary training data, or the models themselves, due to the risk of overall harm. We will approach future work cautiously, along with more research to detect and mitigate instances of misuse of TTS technologies.

Cloud AI Hub

Cloud AI Hub[35] was proposed as a hosted repository of plug-and-play AI components, including end-to-end AI pipelines and out-of-the-box algorithms, making it easy for enterprises and other organizations to share and more readily access a variety of pre-trained machine learning models. Much of the content on AI Hub content would be published by organizations outside of Google, making it easy for the entire community to share ideas, but difficult for us to evaluate all the content along the AI Principles. We assessed the ethical considerations around developing and launching the AI Hub, such as the potential for harmful dual use, abuse, or presentation of misleading information.

In the course of the review, the team developed a strategy to address contributions of potentially risky and harmful content. First, they designed the product to encourage active curation of models by the community with respect to usefulness, unfair bias, and other items. They also developed a plan to support the community by providing resources, tools and assistance (e.g., our inclusive ML guide[36]) to help users identify trustworthy content, and build a reputation engine to help community members build visible expertise. Additionally, we committed to remove problematic content and implement additional machine learning fairness tooling as it becomes available. We also issued service-specific terms for Cloud AI Hub[37].

These safeguards made it more likely that the AI Hub's content ecosystem would be useful and well-maintained and as a result, we proceeded to launch. Future plans for the AI Hub include a public community and marketplace where users will be able to publish both open source and proprietary technology components. If effectively managed, the AI Hub presents an opportunity to promote responsible practices and tools for using and developing AI, including by incorporating educational materials, integrating fairness resources, and promoting content with additional safeguards from trusted developers.

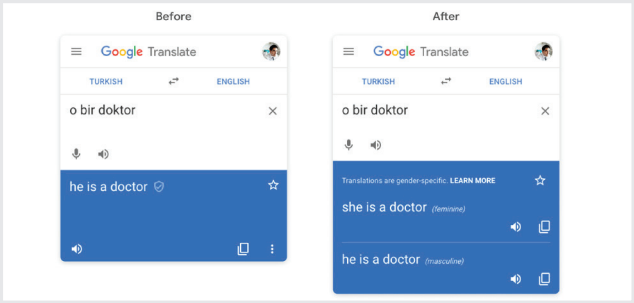
Gender bias in Google Translate

Google Translate learns from hundreds of millions of already-translated examples from the web. Historically, the product has provided only one translation for a query, even if the translation could have either a feminine or masculine form. So when the model produced one translation, it inadvertently replicated gender biases that already existed. For example: it would skew masculine for words like “strong” or “doctor,” and feminine for other words, like “nurse” or “beautiful.” To bring this into alignment with our AI Principles and promote fairness and reduce bias in machine learning, we aimed to address gender bias by providing feminine and masculine translations for some gender-neutral words on the Google Translate website. Notably, we aimed to address bias in this case not by suppressing results but rather by showing multiple valid results.

Implementation of this idea required many steps, including careful data selection of gendered and non-gendered data, development of a system to identify eligibility for gendered translations, design of a new translation system capable of translation into multiple genders, a system to validate results before displaying them to users, and iterative tests and improvements for accuracy and latency.

As a result of these changes, users now see a feminine and masculine translation for a single word—like “surgeon”—when translating from English into French, Italian, Portuguese or Spanish.

Users also get both translations when translating phrases and sentences from Turkish to English. For example, if you type “o bir doktor” in Turkish, you’ll now get “she is a doctor” and “he is a doctor” as the gender-specific translations.



We continue to work on extending gender-specific translations to more languages, and exploring options for addressing non-binary gender in translations. We are also working on addressing similar gender bias issues in other features like query auto-complete.

Community outreach and exchange

We recognize that our efforts in building responsible AI can succeed only with the collaboration and guidance of the wider community, and that it is also important for us to contribute to efforts by other organizations. That is why we regularly engage in conferences to share ideas on the latest progress in the field, and solicit feedback from external experts, policymakers and civil society. These groups inform our perspectives on the potential impacts of AI on society, and enhance our understanding of the wide range of potential benefits and concerns. For example, we held confidential roundtable sessions in the US, UK, France, Belgium, Costa Rica, Korea, Singapore, Ghana, Nigeria and Japan to receive feedback and ideas on early versions of our AI Principles engaging advocacy and business groups, think tanks, and government stakeholders.

To supplement this established variety of dialogue and consultation, we considered an external advisory council as an additional structured way to gather a range of viewpoints on the social and economic impacts of AI. After we announced the plan, however, it became clear that we had not done enough consultation on the types of expertise needed, the role of the council, or how we would take their perspectives into account. While it became clear the council would not be able to function as we had hoped, we remain grateful for the members willing to volunteer their time and expertise, and continue to engage with experts on priority issues and applications.

In the past year, we've discussed broader issues around responsible AI development in sessions with policymakers, civil society groups, and experts in Poland, Germany, Finland, Sweden, Uruguay, Brazil, Mexico, Argentina, Chile, and China, and look forward to continued consultation. Following are some of the community initiatives Google has participated in around responsible development and use of AI.

Community workshops

Credibility on the Web Workshop: Organized workshop for external guests and Google researchers to identify and propose challenges for computer science research on information disorder.

Google AI Ghana Workshop[38]: Hosted inaugural research workshop to strengthen the AI ecosystem in Africa through collaborations with local universities and research centers.

Fairness in ML workshop[39]: Hosted a workshop with attendees from a diverse set of research perspectives to discuss how to build ML algorithms that benefit everyone.

Interpretability workshops: Organized interpretability workshop[40] at ICML (HILL) 2019.

Community Based System Dynamics workshop[41]: Support for Community Based System Dynamics workshop at D4BL 2019.

Black in AI[42]: Participation and funding for workshops and programs supporting Black individuals working in AI.

LatinX in AI[43]: Participation and funding for workshops and programs supporting Latinx individuals working in AI.

WiML Workshops[44] - 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.

Queer in AI workshops[45]: Participation and funding for workshops and social events raising awareness of queer issues in AI and fostering a community of queer researchers.

1st Southeast Asia Machine Learning School[46]: Participation and funding for conference focused on machine learning education.

Abusive Language Workshop[47]: Organized workshop focused on computational detection of abusive language.

ACL Workshop on Gender Bias for Natural Language Processing[48]: Organized workshop and contributed tutorials.

FATE-CV workshop[49]: Googlers co-organized the first workshop in computer vision on Fairness, Accountability, Transparency, and Ethics.

Tutorials: Socially Responsible NLP[50] (WebConf, NAACL Tutorial), Fairness Industry Practices (WWW[51], WSDM[52], KDD), Interpretability at CVPR[53] 2018 and Deep Learning Summer School[54] at University of Toronto/ Vector Institute 2018.

Telepresence at conferences: Pioneering telepresence for underrepresented groups and individuals without visas, including at WiNLP, Black in AI, and NAACL[55].

TechAide in Montreal[56]: Supported a one-day conference to share AI research and raise money for underserved communities in the Montreal area.

International Workshop for Disaster Resilience[57]: Shared work on the Google Flood Forecasting Initiative at this event jointly organized by the National Disaster Management Authority (NDMA) India and the UN Office for Disaster Risk Reduction (UNISDR).

Partnership on AI (PAI): We have participated in numerous workshops and initiatives organised by PAI including on topics such as Fair, Transparent, and Accountable AI; Social and Societal Influences of AI, and Positive Futures from AI.

Competitions and grant programs

AI Impact Challenge[63]**:** 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.

AI4All grant[64]**:** $2M from Google.org to AI4All, a nonprofit working to increase diversity and inclusion in AI, to create a new free digital high school curriculum and learning platform.

Faculty research award program[65]**:** In 2018, this program provided ~$9M in unrestricted gifts for research at institutions around the world. Final proposals were reviewed for alignment with Google's AI Principles and advocated for funding accordingly.

Inclusive Images Competition on Kaggle[66]**:** A challenge to create image recognition systems that can perform well on test images drawn from different geographic distributions than the ones they were trained on. Developed in collaboration with the NeurIPS competition track.

Other engagements

Project Respect[58]**:** an effort to collect positive identity statements from the LGBTIQ+ community into an open dataset for making machine learning models more inclusive.

AI for Social Good Research Network[59]**:** This partnership with the United Nations and the Association of Pacific Rim Universities brings together academics from around the Asia Pacific region to conduct research on ways to build the AI for Good ecosystem.

2019 World Government Summit in Dubai[60]**:** Included a Global Data Commons task force discussion and Global Governance of AI Roundtable, bringing together key actors in the field to share perspectives and identify concrete next steps.

European Commission High-Level Expert Group on AI[61]**:** Participation in the EC initiative to develop policy and investment recommendations for AI, sharing industry and technical expertise.

Singapore Advisory Council on AI and Data[62]**:** Participation in government effort to outline practical measures for organizations using and developing AI, aiming to promote AI adoption while building consumer confidence and trust.

fAIr LAC: Participation and sponsorship support for the Inter-American Development Bank's platform to harness the ethical use of AI for social impact in Latin America and the Caribbean.

OECD Principles on AI: Participation in working group to draft guidelines that promote "AI that is innovative and trustworthy and that respects human rights and democratic values," adopted by OECD member countries along with Argentina, Brazil, Colombia, Costa Rica, Peru, and Romania.

Conclusion

We have learned a lot in the past year, and recognize there is still much more work to do. In the next year we plan to scale our educational efforts; increase the diversity of our teams; build on our technical research and tools; develop and pilot new processes; explore additional forums for external engagement; and continue to share our learnings including through quarterly updates to our Responsible AI Practices[67] and input on wider industry consultations. We are committed to promoting socially valuable applications of AI and advanced technologies, and look forward to continued work with others in the field.

Appendix: Research publications and tools

Research publications

- 50 Years of Test (Un)fairness: Lessons for Machine Learning, Ben Hutchinson, Margaret Mitchell, FAT* 2019.
- A critique of the DeepSec Platform for Security Analysis of Deep Learning Models, Nicholas Carlini, CoRR 2019.
- A General Approach to Adding Differential Privacy to Iterative Training Procedures, H. Brendan McMahan, Galen Andrew, Úlfar Erlingsson, Steve Chien, Ilya Mironov, Nicolas Papernot, Peter Kairouz, NeurIPS 2018.
- A Marauder's Map of Security and Privacy in Machine Learning, Nicolas Papernot, 2018.
- Adversarial Examples Are a Natural Consequence of Test Error in Noise, Nic Ford, Justin Gilmer, Nicolas Carlini, Dogus Cubuk, 2019.
- Adversarial Examples as an Input-Fault Tolerance Problem, Angus Galloway, Anna Golubeva, Graham William Taylor, NeurIPS 2018.
- Adversarial Spheres, Justin Gilmer, Luke Metz, Fartash Faghri, Sam Schoenholz, Maithra Raghu, Martin Wattenberg, Ian Goodfellow, ICLR 2018.
- Adversarially Robust Generalization Requires More Data, Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipra, Kunal Talwar, Aleksander Madry, NeurIPS 2018.
- Amplification by Shuffling: From Local to Central Differential Privacy via Anonymity, Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwa, Abhradeep Thakurta, SODA 2019.
- Anatomy of a Privacy-Safe Large-Scale Information Extraction System Over Email, Ying Sheng, Sandeep Tata, James B. Wendt, Jing Xie, Qi Zhao, Marc Najork, ACM 2018.
- Audio De-identification: A New Entity Recognition Task, Ido Cohn, Itay Laish, Genady Beryozkin, Gang Li, Izhak Shafran, Idan Szpektor, Tzvika Hartman, Avinatan Hassidim, Yossi Matias, NAACL 2019.
- BIM: Towards Quantitative Evaluation of Interpretability Methods with Ground Truth, Mengjiao Yang, Been Kim, 2019.
- BriarPatches: Pixel-Space Interventions for Inducing Demographic Parity, Alexey Alexeevich Gritsenko, Alexander Nicholas D'Amour, James Atwood, Yoni Halpern, D. Sculley, NeurIPS 2018.
- Counterfactual Fairness in Text Classification through Robustness, Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, Alex Beutel, AIES 2019.
- Counterfactual Visual Explanations, Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, Stefan Lee, 2019.
- Deep determinantal generative classifier: robustness on noisy and adversarial samples, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, Jinwoo Shin, ICML 2019.
- Diminishing Returns Shape Constraints for Interpretability and Regularization, Maya Gupta, Dara Bahri, Andrew Cotter, Kevin Canini, NeurIPS 2018.
- Direct Uncertainty Prediction for Medical Second Opinions, Maithra Raghu, Katy Blumer, Rory Abbott Sayres, Ziad Obermeyer, Robert Kleinberg, Sendhil Mullainathan, Jon Kleinberg, ICML 2019.
- Discovering User Bias in Ordinal Voting Systems, Alyssa Whitlock Lees, Chris Welty, SAD-2019.
- Diversity-Sensitive Conditional Generative Adversarial Networks, Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, Honglak Lee, ICLR 2019.
- Do Neural Networks Show Gestalt Phenomena? An Exploration of the Law of Closure, Been Kim, Emily Reif, Martin Wattenberg, Samy Bengio, 2019.
- Ensuring Fairness in Machine Learning to Advance Health Equity, Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, Marshall H. Chin, Ann Intern Med. 2018.
- Explaining Classifiers with Causal Concept Effect (submitted for publication)
- Exploiting Excessive Invariance caused by Norm-Bounded Adversarial Robustness, Jörn-Henrik Jacobsen, Jens Behrmannn, Nicholas Carlini, Florian Tramèr, Nicolas Papernot, ICLR 2019.

- Failure Modes of Variational Inference for Decision Making, Carlos Riquelme, Matthew Johnson, Matt Hoffman, ICML 2018.
- Fairness in Recommendation Ranking through Pairwise Comparisons, Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, Cristos Goodrow}, KDD 2019.
- Fairness Sample Complexity and the Case for Human Intervention, Alyssa Whitlock Lees, Ananth Balashankar, CHI 2019.
- Federated Heavy Hitters with Differential Privacy, Wennan Zhu, Peter Kairouz, Haicheng Sun, Brendan McMahan, Wei Li, 2019.
- Federated Learning for Mobile Keyboard Prediction, Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, Daniel Ramage, 2018.
- Flexibly Fair Representation Learning by Disentanglement, Elliot Creager, David Madras, Jorn Jacobsen, Marissa Weis, Kevin Jordan Swersky, Toniann Pitassi, Richard Zemel, ICML 2019.
- Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy, Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg Corrado, Lily Peng, Dale Webster, Ophthalmology 2018.
- Hiding Images Within Images, Shumeet Baluja, IEEE Transactions on Pattern Analysis and Machine Intelligence 2019.
- Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making, Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, Michael Terry, 2019.
- Human-in-the-Loop Interpretability Prior, Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J. Gershman, Finale Doshi-Velez, NeurIPS 2018.
- Identifying and Correcting Label Bias in Machine Learning, Heinrich Jiang, Ofir Nachum, 2019.
- Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition, Yao Qin, Nicholas Carlini, Ian Goodfellow, Garrison Cottrell, Colin Raffel, ICML 2019.
- Interpreting Black Box Predictions using Fisher Kernels, Rajiv Khanna, Been Kim, Joydeep Ghosh, Oluwasanmi Koyejo, 2018.
- Intriguing Properties of Adversarial Examples, Ekin Dogus Cubuk, Barret Zoph, Sam Schoenholz, Quoc V. Le, ICLR 2018.
- Learning Differentially Private Recurrent Language Models , Brendan McMahan, Daniel Ramage, Kunal Talwar, Li Zhang, ICLR 2018.
- Learning to Attack: Adversarial Transformation Networks, Shumeet Baluja, Ian Fischer, AAAI-2018.
- Leave no Trace: Learning to Reset for Safe and Autonomous Reinforcement Learning, Benjamin Eysenbach, Shane Gu, Julian Ibarz, Sergey Levine, ICLR 2018.
- Metric-optimized Example Weights, Sen Zhao, Mahdi Milani Fard, Harikrishna Narasimhan, Maya Gupta, ICML 2019.
- Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns, Jason Baldridge, Kellie Webster, Marta Recasens, Vera Axelrod, ACL 2018.
- Mitigating Unwanted Biases with Adversarial Learning, Blake Lemoine, Brian Zhang, Margaret Mitchell, 2018.
- Model Cards for Model Reporting, Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru, FAT* 2019.
- Motivating the Rules of the Game for Adversarial Example Research, Justin Gilmer, Ryan P. Adams, Ian Goodfellow, David Andersen, George E. Dahl, 2018. Justin Gilmer, Ryan P. Adams, Ian Goodfellow, David Andersen, George E. Dahl, 2018.
- Multi-Task Learning for Personal Search Ranking with Query Clustering, Jiaming Shen, Maryam Karimzadehgan, Michael Bendersky, Zhen Qin, Don Metzler, CIKM 2018.
- Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification, Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, Lucy Vasserman, 2019.
- On Evaluating Adversarial Robustness, Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, Alexey Kurakin, 2019.
- Optimal Noise-Adding Mechanism in Additive Differential Privacy, Quan Geng, Wei Ding, Ruiqi Guo, Sanjiv Kumar, AISTATS 2019.

- Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals, Andrew Cotter, Heinrich Jiang, Serena Wang, Taman Narayan, Maya Gupta, Seungil You, Karthik Sridharan, 2018.
- Pairwise Fairness for Ranking and Regression, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, Serena Wang, 2019.
- Pareto-Efficient Fairness for Skewed Subgroup Data, Alyssa Whitlock Lees, Ananth Balashankar, Chris Welty, Lakshminarayanan Subramanian, AISG 2019.
- Playing the Game of Universal Adversarial Perturbations, Julien Perolat, Mateusz Malinowski, Bilal Piot, Olivier Pietquin, 2018.
- Privacy Amplification by Iteration, Vitaly Feldman, Ilya Mironov, Abhradeep Thakurta, Kunal Talwar, FOCS 2018.
- Privacy in Geospatial Applications and Location-Based Social Networks, Igor Bilogrevic, Handbook of Mobile Data Privacy 2018.
- Privacy-preserving Prediction, Cynthia Dwork, Vitaly Feldman, COLT 2018.
- Private Selection from Private Candidates, Jingcheng Liu, Kunal Talwar, 2018.
- Proxy Fairness, Maya Gupta, Andrew Cotter, Mahdi Milani Fard, Serena Wang, 2018.
- Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements, Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, Ed H. Chi, AIES 2019.
- Risk-Sensitive Generative Adversarial Imitation Learning, Jonathan Lacotte, Mohammad Ghavamzadeh, Yinlam Chow, Marco Pavone, AISTATS 2018.
- Sanity Checks for Saliency Maps, Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, Been Kim, NeurIPS 2018.
- Scalable Private Learning with PATE, Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, Úlfar Erlingsson, ICRL 2018.
- Shape Constraints for Set Functions, Andrew Cotter, Maya Gupta, Heinrich Jiang, Erez Louidor, James Muller, Tamann Narayan, Serena Wang, Tao Zhu, ICML 2019.
- Text Embeddings Contain Bias. Here's Why That Matters, Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, Úlfar Erlingsson, ICLR 2018.
- The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks, Nicholas Carlini, Chang Liu, Jernej Kos, Úlfar Erlingsson, Dawn Song, USENIX Security 2019.
- The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets, Nicholas Carlini, Chang Liu, Jernej Kos, Úlfar Erlingsson, Dawn Song, 2018.
- To Trust or Not To Trust a Classifier, Heinrich Jiang, Been Kim, Melody Y. Guan, Maya Gupta, NeurIPS 2018.
- Tough Times at Transitional Homeless Shelters: Considering the Impact of Financial Insecurity on Digital Security and Privacy, Manya Sleeper, Tara Matthews, Kathleen O'Leary, Anna Turner, Jill Palzkill Woelfer, Martin Shelton, Andrew Oplinger, Andreas Schou, Sunny Consolvo, CHI 2019.
- Towards Automatic Concept-based Explanations, Amirata Ghorbani, James Wexler, James Zou, Been Kim, 2019.
- Towards Equitable AI for the Next Billion Users, Nithya Sambasivan, Jess Scon Holbrook, ACM 2019.
- Towards Federated Learning at Scale: System Design, Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, H. Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, Jason Roselander, SysML 2019.
- Training On-Device Ranking Models from Cross-User Interactions in a Privacy-Preserving Fashion, Marc Najork, DESIRES 2018.
- Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints, Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, Seungil You, 2018.
- Two Player Games for Efficient Non-Convex Constrained Optimization, Andrew Cotter, Heinrich Jiang, Karthik Sridharan, PMLR 2019.

- Understanding and correcting pathologies in the training of learned optimizers, Luke Metz, Niru Maheswaranathan, Jeremy Nixon, Daniel Freeman, Jascha Sohl-dickstein, ICML 2019.
- Unrestricted Adversarial Examples, Tom B. Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, Ian Goodfellow, 2018.
- Visualizing and Measuring the Geometry of BERT, Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, Martin Wattenberg, 2019.
- What made you do this? Understanding black-box decisions with sufficient input subsets, Brandon Carter, Jonas Mueller, Siddhartha Jain, David Gifford, PMLR 2019.

Tools

- Creatability: A set of experiments made in collaboration with creators and allies in the accessibility community, exploring how creative tools can be made more accessible using web and AI technology.
- Critiquing Protein Family Classification Models Using Sufficient Input Subsets: A local explanation framework for interpreting black-box functions to find a sufficient input subset that is a minimal set of input features whose observed values alone suffice for the same decision to be reached.
- Dataset Search: A new search tool for anyone to find useful and interesting datasets on the Web, from various scientific disciplines, such as environmental or social science, to government or economic data from around the world.
- Machine Learning Crash Course fairness module: Developed an Intro to Fairness module to our machine learning crash course, available for anyone to take for free.
- Open Images eXtended: A new branch of Google's Open Images dataset, to expand the diversity of cultures and people represented in imagery training data. Includes a prototype data card, providing transparency on the background, characteristics, and composition of the dataset.
- People + AI Guidebook: A toolkit of methods and decision-making frameworks for teams to build human-centered AI products. Includes contributions from 40 Google product teams.
- Project Respect: an effort to collect positive identity statements from the LGBTIQ+ community into an open dataset for making machine learning models more inclusive.
- TensorFlow Federated: Federated learning is a distributed machine learning approach in which a fleet of devices coordinates to train a shared global model from locally-stored data; open-sourced as TensorFlow Federated.
- TensorFlow Privacy: An open source library for developers to train ML models with differential privacy, and for researchers to advance the state of the art in ML with strong privacy guarantees.
- TF Constrained Optimization Library: A library for optimizing inequality-constrained problems in TensorFlow.
- Threshold agnostic metrics tutorial: Tutorial exploring bias AUC metrics.
- What-If Tool: An interactive tool built into TensorBoard, and usable in Jupyter and Colaboratory notebooks, to analyze model performance on a dataset and observe the effects of changes to input data or model constraints; enables users to visualize biases and the effects of various fairness constraints as well as compare performance across multiple models.

End notes

1 https://www.blog.google/products/translate/reducing-gender-bias-google-translate/

2 https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html

3 https://blog.google/technology/safety-security/privacy-everyone-io/

4 https://ai.google/responsibilities/responsible-ai-practices/

5 https://www.blog.google/technology/ai/google-ai-principles-updates-six-months/

6 https://ai.googleblog.com/2018/12/adding-diversity-to-images-with-open.html

7 https://ai.google/static/documents/datasets/open-images-extended-crowdsourced.pdf

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

9 https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html

10 https://colab.sandbox.google.com/github/tensorflow/tensorboard/blob/master/tensorboard/plugins/interactive_inference/What_If_Tool_Notebook_Usage.ipynb

11 https://annals.org/aim/article-abstract/2717119/ensuring-fairness-machine-learning-advance-health-equity

12 https://arxiv.org/abs/1903.04561

13 https://github.com/conversationai/unintended-ml-bias-analysis/blob/master/presentations/FAT_star_tutorial.md

14 https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview

15 https://ai.google/research/pubs/pub48107

16 https://arxiv.org/pdf/1809.10610.pdf

17 https://arxiv.org/abs/1811.10104

18 https://developers.google.com/machine-learning/crash-course/fairness/video-lecture

19 https://github.com/tensorflow/tcav/

20 https://arxiv.org/abs/1902.02960

21 https://en.unesco.org/news/learning-digital-age-smart-cities-among-innovations-taking-centre-stage-unesco-netexplo-forum

22 https://arxiv.org/abs/1810.10118

23 https://arxiv.org/abs/1805.11783

24 https://github.com/google/TrustScore

25 https://arxiv.org/abs/1805.11571

26 https://federated.withgoogle.com/

27 https://arxiv.org/abs/1602.05629

28 https://arxiv.org/pdf/1811.03604.pdf

29 https://www.tensorflow.org/federated

30 https://medium.com/tensorflow/introducing-tensorflow-privacy-learning-with-differential-privacy-for-training-data-b143c5e801b6

31 https://arxiv.org/abs/1902.06705

32 https://pair.withgoogle.com/

33 https://experiments.withgoogle.com/collection/creatability

34 https://www.blog.google/products/search/making-it-easier-discover-datasets/

35 https://cloud.google.com/ai-hub/

36 https://cloud.google.com/inclusive-ml/

37 https://aihub.cloud.google.com/u/0/aihub-tos

38 https://sites.google.com/corp/view/aisummit-accra/home

39 https://sites.google.com/corp/view/mlfairnessworkshop/home?authuser=0

40 https://sites.google.com/corp/view/hill2019/home

41 https://accelerate.withgoogle.com/stories/exploring-systems-dynamics-inclusive-ml-and-societal-impact-meet-googlers-donald-martin-and-jamaal-sebastian-barnes

42 https://blackinai.github.io/

43 https://www.latinxinai.org/

44 https://wimlworkshop.org/

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

46 https://www.sea-mls.com/home

47 https://sites.google.com/corp/view/alw3/home?authuser=0

48 https://genderbiasnlp.talp.cat/

49 https://sites.google.com/corp/view/fatecv

50 https://sites.google.com/corp/view/srnlp/home

51 http://webconf/

52 https://www.slideshare.net/KrishnaramKenthapadi/fairnessaware-machine-learning-practical-challenges-and-lessons-learned-wsdm-2019-tutorial

53 https://interpretablevision.github.io/index_cvpr2018.html

54 https://dlrlsummerschool.ca/2018-event/

55 https://naacl2019.org/blog/remote-presentations/

56 https://www.techaidemontreal.org/ai-conference

57 https://resilientinfra.org/iwdri/about.php

58 https://projectrespect.withgoogle.com/

59 https://www.unescap.org/news/google-and-united-nations-economic-and-social-commission-asia-and-pacific-host-asia-pacific-ai

60 https://www.worldgovernmentsummit.org/docs/default-source/publication/2019/wgs_agenda_2019_ar---en.pdf

61 https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

62 https://www.gov.sg/resources/sgpc/media_releases/imda/press_release/P-20181130-1?authkey=f5e3c0cf-1284-433d-8a5d-77c33913f09a

63 https://ai.google/social-good/impact-challenge/

64 https://blog.google/outreach-initiatives/google-org/ai4all-participants-tell-allsummer-camps-get-girls-involved-ai-and-tech/

65 https://ai.google/research/outreach/faculty-research-awards/

66 https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

67 https://ai.google/responsibilities/responsible-ai-practices/

Google