

Google

Responsible
Development of AI

Introduction

Scientific progress is putting the challenges and complexity of the world's problems into starker relief. Fortunately, it is also yielding new technological tools to help us in tackling these problems - most notably AI. The pace of AI research breakthroughs is now being matched by real world application, offering new possibilities for boosting productivity and insight across virtually every field. At the same time, AI is shining new light on (and sometimes magnifying) old and difficult questions. For example, how do we, as a society, think about fairness? About building inclusive experiences? About equipping the workforce for the jobs of the future?

Harnessed appropriately, we believe AI can deliver great benefits for economies and society, and support decision-making which is fairer, safer and more inclusive and informed. But such promise will not be realized without great care and effort, including confronting the risks of AI being misused and taking steps to minimize.

AI's influence on the world will be determined by the choices people make in embracing it. Just as musicians choose which instruments and music to play for specific audiences and purposes, and within venue constraints, so too do programmers and businesses choose which techniques to apply and to what ends, within the boundaries set by governments and cultural acceptance. Ultimately it's up to countries and societies to choose how they want to harness the benefits of AI, and to establish the right frameworks for their development. Google is committed to engaging where we can be helpful, and hope that this document contributes to that discussion.

Overview of Google's approach

As one of the leaders in the field, we acknowledge that Google has an obligation to develop and apply AI thoughtfully and responsibly, and to support others to do the same. Like others in the industry we have outlined general principles that are important to us¹ (see box), and we are committed to establishing processes and governance structures to help us adhere to the principles in spirit and substance.

In practice, the manner in which the AI principles can be delivered is often dependent on the technological possibilities. Google is at the forefront of research to help expand the scope of what is possible for ourselves and others, including tools and guidance for developers (see box on the next page).

A good example of this is the People and AI research group (PAIR) within Google's research team, dedicated to helping to tackle issues such as concerns around algorithmic bias, explainability and usability. They create open-source tools to help developers better understand the risks of their systems and identify and fix problems that may exist (E.g., dataset visualization tools like Facets). More generally, they seek to educate and promote the state of the art by publishing research and sharing it with the wider community.

Google AI principles

- **AI should be socially beneficial:** with the likely benefit to people and society substantially exceeding the foreseeable risks and downsides
- **AI should not create or reinforce 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
- **AI should 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
- **AI should be accountable to people:** providing appropriate opportunities for feedback, relevant explanations and appeal, and subject to appropriate human direction and control
- **AI should incorporate privacy design principles:** encouraging architectures with privacy safeguards, and providing appropriate transparency and control over the use of data
- **AI development should uphold high standards of scientific excellence:** Technological innovation is rooted in the scientific method and a commitment to open inquiry, intellectual rigor, integrity, and collaboration
- **AI should be made available for uses that accord with these principles:** We will work to limit potentially harmful or abusive applications

Some highlights of our work



Interpretability

Google is working intensely to advance the areas of AI interpretability and accountability, through open-sourcing tools and publishing research. For instance:

- **Tensor Flow Lattice:** enabling anyone to train flexible models that capture a priori knowledge about whether an input should only have a positive effect on an output²
- **Tensor Flow Debugger:** enabling developers to look inside models during training³
- **Building blocks of interpretability:** illustrating how different techniques can be combined to provide powerful interfaces for explaining neural network outputs⁴



Privacy

Google has long supported efforts in the research and development of privacy and anonymity techniques for AI systems, including publishing new open-source code as privacy-protection best practices. For example:

- **Our open-source RAPPOR technology:** deployed worldwide as the first large-scale data-collection mechanism with strong protection guarantees of local differential privacy⁸
- **Secure aggregation protocol for federated learning model updates:** provides strong cryptographic privacy for individual user's updates, averaging only updates from large groups of participants⁹



Security

One of the biggest threats to AI systems currently comes from “adversarial attacks”, in which bad actors fool the system by making very small, not human detectable, changes to model inputs. Fortunately such attacks are very difficult and hence not (yet) widespread, but Google researchers are at the forefront of tackling them. Publicly released research includes:

- **Adversarial Logit Pairing (ALP):** state of the art in defenses against adversarial examples⁵
- **Ensemble Adversarial Training:** the previous state of the art (before ALP) in defenses against black-box adversarial examples, developed in collaboration with Stanford⁶
- **CleverHans:** a machine learning security research library maintained by Google team⁷



Fairness

AI systems are shaped by what their training data leaves out and what it over-represents. Human biases within the data, model design, and methods for training and testing can lead to outcomes that affect different groups of people differently. Addressing these disparate outcomes is a primary goal in the emerging research area of fairness in machine learning. Google is an active contributor to this field, including in the provision of developer tools. For example:

- **Facets:** interactive visualization tool that lets developers see a holistic picture of their training data at different granularities¹⁰
- **Mmd-critic:** exploratory data analysis tool that looks for statistical minorities in the data¹¹

In addition to these principles, we are also committing to not pursue some applications. In particular, we will not design or deploy AI in weapons or other technologies designed to cause or directly facilitate injury to people; or in technologies that gather or use information for surveillance violating internationally accepted norms; or technologies for any purpose that contravene widely accepted principles of international law and human rights. More generally, for any AI applications 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.

We will incorporate these principles and restrictions into our development work and our review of implementations and commercial agreements. We will use a combination of central teams reviewing issues like privacy, discrimination, and legal compliance, as well as decentralized assessments by each Product Area. A crossfunctional AI Review Council will be available to assess challenging issues. We will also integrate various oversight mechanisms and working practices into our existing operating processes. For example:

- The internal launch process which every new product or significant feature undergoes will include a check for fit against topline principles.
- During 2017 Google's Trust and Safety team piloted an initiative to provide product teams with expert help to assess risks and test for possible bias, which is now being rolled out company-wide. Support includes templates, tips, case studies, provision of diverse 'dogfooding' test groups, and hands-on help in experiment design.
- Our Privacy Working Group and ML Fairness teams will assess relevant issues in the context of new tools that incorporate AI.
- Each Product Area has multiple product counsel assigned to review new product launches for international legal compliance.
- We are experimenting with various mechanisms to make developers more aware of dataset limitations, so they can better select the right dataset for their application. One possibility, for instance, would be to include consistent information about datasets, akin to nutrition labels on food packaging. For similar reasons we are exploring providing more detailed guidance on the appropriate use of pre-trained models that we make available to developers.
- We recognize that ethical considerations need to be thought about at every stage of product development. We are experimenting with various ways to help our teams do this, such as modules added to our internal machine learning training (E.g., on algorithmic bias) and more formal collaborations with universities to develop ethics case studies and custom courses.

While we strive to make AI accessible to everyone (see box), it is worth noting that many of the toughest internal debates so far have related to concerns about what users of Google's AI tools might do, applying them in ways we did not foresee or condone.

When considering selling or distributing technologies that could foreseeably be misused, we take into account a number of factors including: whether the technology in question is generally available or unique to Google; how readily adaptable it could be to a harmful use, and potential magnitude of impact; and likelihood of risk linked to our level of involvement.

Finally, we recognize that this is not something Google can or should seek to solve alone. It is vital that the discussion about the responsible development and application of AI involves a broad range of stakeholders and perspectives. To facilitate this conversation we regularly engage with external experts through academic and industry conferences, as well as in policy forums. Google also co-founded the Partnership on AI, which was established to study and formulate best practices on AI technologies, to advance the public's understanding of AI, and to serve as an open platform for discussion and engagement about AI and its influences on people and society.

Google is helping everyone access AI

AI has flourished in part because of a set of common norms that encourage research results to be published and shared openly. Google is committed to preserving these community principles through publishing our research results and actively participating in conferences.

We also release open source tools for researchers and other experts to use. For example:

- We open sourced TensorFlow - Google's internal machine learning toolkit to allow anyone to experiment in the space and advance the state of the art
- We have invested in creating and sharing large datasets to support machine learning researchers for many data types, including speech commands, photos and video, online discussion, audio effects, and crowdsourced drawings
- The Learn With Google AI website offers free lessons, tutorials and hands-on exercises for people at all experience levels seeking to learn to use AI

Finally, Google Cloud lowers the barrier to entry and makes AI available to the largest possible community of developers, researchers and businesses. Our platform offers modern machine learning services built upon pre-trained models that can bring unmatched scale and speed to business applications.

How policy makers can help

Like any technology, there is nothing predestined about the impact of AI. While AI researchers can lay the groundwork for what's technically feasible, AI's impact in practice will depend on the appetite expressed for it by industry and society, and the guidance and boundaries set for its application by government. Policy makers thus are crucial to crafting the vision and establishing the frameworks that will underpin AI's development.

While the specifics of what is relevant and practicable will vary by country, there are some common themes worth exploring by policy makers who seek to champion the responsible use of AI. Providing reassurance, support and encouragement to the general public and businesses alike about AI is crucial, as is clarifying governance frameworks. Making government a role model for responsibly embracing AI would reap benefits directly as well as be a best practice guide for others. Finally, there are some particular challenges associated with AI relating to data access and research funding that policy makers are well-placed to help overcome.

More detail on each of these themes follows, along with some thought-starter suggestions of tactics for addressing them based on examples from around the world. We hope this will serve as a useful spark for policy ideas, and look forward to greater engagement around AI development.

Actions to encourage responsible use of AI

1. Help to boost public confidence and understanding of AI
2. Stimulate uptake of AI within priority industry sectors
3. Facilitate AI research that helps tackle barriers to implementation
4. Encourage responsible data sharing to boost data available for training AI systems
5. Promote constructive governance frameworks and build expertise in government bodies
6. Make government a role model for responsibly embracing AI
7. Take steps to prepare for workforce transition

1. Help to boost public confidence and understanding of AI

AI offers great opportunity to benefit society, by providing crucial help in advancing science, improving access to medical care, and boosting economic productivity. But realizing such promise is reliant on societal acceptance. Currently, public confidence in AI is undermined by science fiction visions which naturally dominate headlines, and distract from less vivid but more immediate dangers such as the risk of bias and malicious use. Fears of AI are also often conflated with general concerns about the future shape of work and rising inequality. Governments have a key role to play alongside industry in acknowledging and addressing all such concerns, and providing a balanced, facts-based picture of the opportunities and challenges which reflect views across disciplines and walks of life. In areas where public trust is especially low, it may also be reassuring to set up expert advisory forums to provide a framework for more formal ongoing engagement and monitoring.

Activities to consider

- Hold forums to gather and synthesize views from cross-sections of society
- Hold expert sessions open to the public to facilitate more grounded and informed debate on key topics
- Stage a series of citizen juries to debate key AI issues and provide recommendations
- Stage formal inquiries to gather advice and demonstrate attentiveness to public input
- Establish expert advisory committee(s) to provide a locus for public engagement
- Enlist support of national science agencies and NGOs in creative public engagement—E.g., celebrity scientist lectures, crossover arts/science events
- Run education campaigns that highlight AI's application to everyday life—E.g., in public sector services, healthcare—to attract and connect to the general public, not just experts
- Offer training grants to encourage people from diverse backgrounds to learn about AI, in order to bring fresh perspectives and allow for wider community representation

2. Stimulate uptake of AI within priority sectors

AI's economic promise will only become reality if it is applied in a meaningful way by industry. Doing so requires a thorough understanding of the kinds of problems that AI is good at tackling, current or inherent limitations, and the resources (tools, data, expertise, computing power) needed to implement AI solutions. Governments can act as a useful signpost in nudging businesses to explore and invest in AI opportunities. Another important lever is government backing to facilitate training in applied AI, and dissemination of best practice and standards.

Activities to consider

- Run studies/surveys to assess the industry's awareness of AI, interest in using it and barriers preventing its uptake
- Appoint expert group to advise on priorities and champion AI to businesses, potentially even to the extent of providing seed capital/expert resources akin to an AI incubator
- Provide incentives to businesses who embrace AI to catalyze their interest
- Offer subsidies to support investment in the physical infrastructure underpinning AI in regions where it is lacking - E.g., discounts on cost of electricity, faster capex depreciation
- Provide funding to support AI application in areas of great social need - E.g., crisis response
- Encourage researchers and businesses to create and share datasets relevant to key sectors, while respecting privacy norms
- Look for tangible ways to make it easier for businesses to access AI tools, including through cloud-based services (E.g., by providing more flexible rules around data localization)
- Provide incentives to catalyze more cross-faculty collaboration between computer science and other industry-focused academic fields (E.g., CS +agriculture, CS+medicine)
- Encourage universities to include training on applying AI across their curriculum (not only in engineering), so the next generation of graduates to enter industry are well-equipped
- Grants to support the development and provision of AI-oriented vocational training for people employed or seeking jobs in priority sectors

3. Facilitate AI research which helps tackle barriers to implementation

Despite recent progress, many research challenges remain to be addressed before AI can realise its full potential. For example, AI systems need to become more explainable, more efficient in terms of the scale of data and computation used to train models, and easier for more people to use and build. To have a thriving AI ecosystem it is important that such fundamental, basic research is not driven solely by the private sector; new high-risk, high-reward research areas may be possible only through public funding. There may also be applications of wider societal benefit that academia and government-funded research organisations are the most natural fit to tackle.

Activities to consider

- Earmark funding for AI research and advanced study at national institutions
- Establish local centers of excellence for AI research and applications
- Open access to publicly-funded research and the resulting datasets
- Perform AI research in key areas of national importance at government labs and agencies
- Create frameworks to foster public / private sector collaboration on AI research

4. Encourage responsible data sharing to boost data available for training AI systems

Machine learning models need datasets for their training, carefully curated and tailored to best represent and address the problem being tackled. Governments could help to boost R&D in AI by creating a framework that incentivizes and makes it easier to create, share and re-use datasets relevant to priority fields of application, in a manner that respects user expectations of privacy. In parallel, added clarity is always welcome in the practical interpretation of data regulations in terms of the difficult trade-offs they embody between societal benefit and individual rights. (E.g., in Europe under GDPR, what counts as a scientific purpose? Outside of Europe, are there any circumstances in which data minimization may not be the most important principle to uphold?)

Activities to consider

- Establish standard terms and mechanisms for privacy-friendly data-sharing, to help reduce the legal and administrative burden of negotiating such transactions
- Make available more public datasets, especially in priority subject realms for innovation
- Provide incentives for researchers who receive public funding to publish datasets associated with their research in machine-readable format, while still respecting privacy norms
- Look for ways to make it easier for people to export and share personal data, if they wish to contribute it for research or use in other applications
- Set up an expert body able to provide timely input and guide researchers in assessing tradeoffs in societal benefits vs individual rights

5. Promote constructive governance frameworks and build expertise within oversight bodies

AI can impact society in a variety of ways, which is why government has such an important role to play alongside industry to ensure good outcomes. A number of existing sectoral regulations, from health care to transportation to communications, already govern AI implementations. Sectoral experts will typically be the best placed to assess context-specific uses, assessing the impact and results of new technologies, but may need support to build AI expertise. As AI advances, governments should expand their technological expertise and explore various cooperative frameworks to minimize issues and maximize AI's potential. Consensus-driven best practices and self-regulatory bodies can also contribute to creating flexible and nuanced approaches.

Activities to consider

- Commission a sector-by-sector analysis of how existing regulatory schemes apply to AI-enabled systems, and what gaps (if any) exist
- Identify any existing constraints which hamper responsible use of AI and seek a solution
 - E.g., inferring race is essential to check that systems aren't racially biased, but existing laws around discrimination and privacy can make this problematic
 - E.g., on-device AI has different risks and characteristics from cloud AI, and the nuances of this may not be reflected by "one size fits all" data protection rules
 - E.g., copyright rules can restrict data available for use in training AI systems, which may undermine efforts at reducing bias if data from key segments are excluded
- Funding to boost in-house technical expertise at regulators in sectors facing greatest potential for disruption
- Appoint advisory committee or lead POC to coordinate on AI governance issues, including representation from the AI research community, industry, and civil society
- Engage with governance bodies around the world to share and learn from experiences
- Encourage industry to share best practices and promote codes of conduct
- Ethics training for government-funded researchers (analogous to research ethics training required for bioscience researchers funded by the US NIH)

6. Make government a role model for responsibly embracing AI

AI's potential to boost productivity and service quality is just as applicable to the public sector as to business. Government can thus lead in showcasing best practice for prioritizing opportunities to embrace AI, as well as demonstrating how it can practically and sensitively be applied. More generally, governments also have a valuable role to play as a catalyst to progress by making available public data sets, which others can use to develop services (akin to the way that the Imagenet database fueled advances in computer vision).

Activities to consider

- Appoint expert group to provide pragmatic actionable guidance on AI application
- Develop basic principles-based guidance for government agencies to use in funding projects that have an AI component or in systems acquisitions. Incorporate in request for proposal and oversight of activities
- Encourage pilot programs to accelerate the use of AI in improving citizen services
- Look for opportunities to boost internal AI expertise, by hiring or partnering with experts
- Partner with platforms like Kaggle to run competitions that use public datasets to help improve public service offerings
- Identify barriers and invest in infrastructure and training to encourage government bodies to improve the curation and sharing of new public datasets

7. Take steps to prepare for workforce transition

There is general consensus that AI will bring about some reconfiguration of employment, even if the pace and scale of impact is as yet unknown. Governments can play a key role in empowering people to develop skills for the future, such as through programs to improve digital literacy in schools or providing incentives to boost work-based learning. To make sure these skills can be put to good use, it is important to improve flexibility and mobility in the labor market - making it easier for businesses to hire, and removing barriers to employment such as excessive credentialing or licensing. In parallel, initiatives to bolster entrepreneurship and foster resilience could help to better equip people to spot and grasp opportunities as they arise. Finally, it is important to reflect on social safety nets and consider if they need to evolve in terms of funding or provision in light of the changing employment landscape.

Activities to consider

- Partner with industry in priority sectors to establish next generation apprenticeship schemes, particularly designed for experienced workers facing transition
- Liaise with employers and employee representatives to review occupational licensing and ensure restrictions are still justified (E.g., on grounds of public safety)
- Establish regional expert groups (with academic, business and labour representatives) who can help to set vocational training priorities based on local employment trends
- Provide support and incentives to workers (and their families) who are willing to relocate, to help businesses fill vacancies in priority sectors and geographies
- Offer bursaries or tax-advantaged savings accounts to all adults to encourage them to engage in lifelong learning
- Fund research and promote pilot schemes to help identify and disseminate best practices for on-the-job training - including experimenting with using AI to tailor and deliver just-in-time education
- Boost pathways for jobseekers to less formal (but proven) learning opportunities such as targeted certification programs or technology “boot camps”
- Research and experiment with new approaches to social safety nets

End notes

- 1 A longer description of Google's AI Principles is viewable at <https://ai.google/principles>
- 2 TensorFlow Lattice (open source) enables training models that capture prior knowledge about whether an input can or should only have a monotonic effect on an output. For example, the input "time since last cleaned the bathroom" should only have a positive impact on the predicted probability that "it's time to clean that bathroom again". More information at <https://bit.ly/2ygocGD>
- 3 TensorFlow Debugger (open source) is a TensorFlow package enables developers to look inside models during training. More details are at <https://www.tensorflow.org/guide/debugger>
- 4 More information about our work on building blocks of interpretability is viewable online at <https://distill.pub/2018/building-blocks/>
- 5 More details of our work is available in the research paper "Adversarial Logit Pairing" by Kannan H et al (2018). Viewable online at <https://arxiv.org/abs/1803.06373>
- 6 For more information see "Ensemble Adversarial Training: Attacks and Defenses" by Tramèr F et al (2018). Viewable online at <https://openreview.net/forum?id=rkZvSe-RZ>
- 7 The CleverHans library is at <https://github.com/tensorflow/cleverhans>
- 8 RAPPOR (open source) is a mechanism for large-scale data collection with strong guarantees of local differential privacy protection. More information is available in this 2014 blogpost <https://bit.ly/2RehPPz>
- 9 For more information about our secure aggregation protocol work see "Practical Secure Aggregation for Privacy Preserving Machine Learning" by Bonawitz K et al (2017). Viewable online at <https://eprint.iacr.org/2017/281>
- 10 Facets (open source) offers a tool for visualizing, analyzing and understanding dataset composition. See more at <https://pair-code.github.io/facets/>
- 11 More information on Mmd-critic is in the research paper "Examples are not Enough, Learn to Criticize! Criticism for Interpretability" by Kim B et al (2016). Viewable online at <https://bit.ly/2GVF1NJ>

Google