Google

Perspectives on Issues in AI Governance

Table of contents

Overview

We have long since progressed beyond an era when advances in AI research were confined to the lab. AI has now become a real-world application technology and part of the fabric of modern life. 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, which includes consideration of how its development and usage should be governed, and what degree of legal and ethical oversight — by whom, and when — is needed.

To date, self- and co-regulatory approaches informed by current laws and perspectives from companies, academia, and associated technical bodies have been largely successful at curbing inopportune AI use. We believe in the vast majority of instances such approaches will continue to suffice, within the constraints provided by existing governance mechanisms (e.g., sector-specific regulatory bodies).

However, this does not mean that there is no need for action by government. To the contrary, this paper is a call for governments and civil society groups worldwide to make a substantive contribution to the AI governance discussion.

Specifically, we highlight five areas where government, in collaboration with wider civil society and AI practitioners, has a crucial role to play in clarifying expectations about AI's application on a context-specific basis. These include explainability standards, approaches to appraising fairness, safety considerations, requirements for human-AI collaboration, and general liability frameworks.

For each area we have provided commentary on the issues and suggestions of concrete actions that government, supported by other stakeholders, could take to provide greater guidance. These suggestions are summarized in **Box 1** and represent practical things that we believe would make a demonstrable impact in helping to ensure the responsible use of AI.

In this white paper we share our point of view on these concrete issues. Google does not have all the answers; on the contrary, it is crucial for policy stakeholders worldwide to engage in the conversation. As AI technology evolves and our own experience with it grows, we expect that the global community as a whole will continue to learn and additional nuances will emerge, including a fuller understanding of the trade-offs and potential unintended consequences that difficult choices entail.

Our observation is that so far much of the current AI governance debate among policymakers has been high level; we hope this paper can help in evolving the discussion to address pragmatic policy ideas and implementation. The 'rules of the road' for AI (be they in the form of laws or norms) will need to evolve over time to reflect thoughtful and informed consideration of economic and social priorities and attitudes, as well as keeping pace with what is possible technologically.

BOX 1

Key areas for clarification and suggested actions

Explainability standards

- Assemble a collection of best practice explanations along with commentary on their praiseworthy characteristics to provide practical inspiration.
- Provide guidelines for hypothetical use cases so industry can calibrate how to balance the benefits of using complex AI systems against the practical constraints that different standards of explainability impose.
- Describe minimum acceptable standards in different industry sectors and application contexts.

Fairness appraisal

- Articulate frameworks to balance competing goals and definitions of fairness.
- Clarify the relative prioritization of competing factors in some common hypothetical situations, even if this will likely differ across cultures and geographies.

Safety considerations

- Outline basic workflows and standards of documentation for specific application contexts that are sufficient to show due diligence in carrying out safety checks.
- Establish safety certification marks to signify that a service has been assessed as passing specified tests for critical applications.

Human-AI collaboration

- Determine contexts when decision-making should not be fully automated by an AI system, but rather would require a meaningful "human in the loop".
- Assess different approaches to enabling human review and supervision of AI systems.

Liability frameworks

- Evaluate potential weaknesses in existing liability rules and explore complementary rules for specific high-risk applications.
- Consider sector-specific safe harbor frameworks and liability caps in domains where there is a worry that liability laws may otherwise discourage societally beneficial innovation.
- Explore insurance alternatives for settings in which traditional liability rules are inadequate or unworkable.

Background

AI is a powerful, multi-purpose technology with the potential to transform industrial and societal processes alike. Governments thus have an important role to play in collaboration with industry and other stakeholders to ensure good outcomes. While AI researchers, developers, and industry can lay the groundwork for what is technically feasible, it is ultimately up to government and civil society to determine the frameworks within which AI systems are developed and deployed.

It is important to note that this effort is not starting from scratch. There are already many sectoral regulations and legal codes that are broad enough to apply to AI, and established judicial processes for resolving disputes. For instance, AI applications relating to healthcare fall within the remit of medical and health regulators, and are bound by existing rules associated with medical devices, research ethics, and the like. When integrated into physical products or services, AI systems are covered by existing rules associated with product liability and negligence. Human rights laws, such as those relating to privacy and equality, can serve as a starting point in addressing disputes. And of course there are a myriad of other general laws relating to copyright, telecommunications, and so on that are technology-neutral in their framing and thus apply to AI applications.

Given the early stage of AI development, it is important to focus on laws and norms that retain flexibility as new possibilities and problems emerge. This is particularly crucial given that AI, like many technologies, is multi-purpose in nature.

Overall we are confident that existing governance structures will prove to be sufficient in the vast majority of instances. In the rare cases where they are not, we believe that sectoral experts in industry and academia together with practitioners at the forefront of AI application are largely well placed to help identify emerging risks and take steps to mitigate them, in consultation with civil society and government. This multi-stakeholder collaborative approach will allow for the most timely and effective response to concerns about AI without impeding its promise.

However, there are key questions which merit additional oversight and guidance from governments. Setting international standards and norms would relieve pressure on individual countries and regions to advance a controversial use of technology just because others might be doing so, preventing a race to the bottom. While international treaties cannot in themselves prevent violations, they clarify shared expectations of behavior and thus serve as a metric against which sanctions can be imposed for misuse. Such rules would also acknowledge that the impact of AI transcends borders, setting a level playing field within industry and raising the bar for responsible use.

In thinking through these issues, it may be helpful to review how the world has responded to the emergence of other technologies presenting ethical (and at the extreme, existential) questions (see **Box 2**). While there are some similarities, there is no directly comparable technology to AI in terms of the breadth of its application and ease of accessibility. From a governance perspective, AI thus poses a unique challenge both in formulating and enforcing regulations and norms.

BOX 2

Responses to ethical questions raised by other transformative technologies

As with other transformative technologies, AI presents many opportunities to solve important problems and unlock societal and economic value, while also raising new ethical questions. Some examples of how the world has responded to questions raised by earlier technologies include:

Genetic engineering – Concerns around synthetic biology and human germ-line editing were first raised in the 1970s by researchers. This led to a voluntary agreement at the groundbreaking Asilomar gathering to impose self-regulatory restrictions on experiments involving recombinant DNA. The global community of AI researchers was inspired to use a similar approach for their own gatherings — most recently leading to the Asilomar AI principles adopted in 2017.

In vitro fertilization (IVF) and human embryo research – The successful demonstration of human IVF in 1978 offered hope for people struggling to conceive. It also led to governmental restrictions barring research on embryos more than 14 days after fertilization (the "14-day rule"), first in the US and over subsequent years in the UK, Europe, Japan, India and elsewhere. This is an example of national governments taking action independently but in a collaborative way that provided common norms across much of the world.

Nuclear technology – Reactions of atomic nuclei can be used for many beneficial applications, including medical imaging, radiation therapy, smoke detectors, and renewable energy production. Nuclear reactions can also be used to produce highly destructive weapons. While new nations continue to develop nuclear technology, national guidelines and international non-proliferation agreements have proven a strong framework for setting and maintaining expectations of responsible behavior.

Polychlorinated biphenyls (PCBs) – PCBs were first manufactured in the 1920s, with many applications including in coolants, plastics, pesticides, and adhesives. These molecules were later found to be environmental pollutants with considerable toxicity. Production was banned by the US in 1978 and by the Stockholm Convention on Persistent Organic Pollutants in 2001. Many chemicals beyond PCBs, including drugs and explosives, have similar dual-use potential. To promote responsible use of chemistry, chemical practitioners around the world worked together to create The Hague Ethical Guidelines in 2015, which were endorsed by the International Union of Pure and Applied Chemistry (IUPAC) and used to develop the American Chemical Society's Global Chemist Code of Ethics in 2016.

Space exploration – The goal of exploring the larger universe captured public imagination while catalyzing many science and engineering breakthroughs. Arguably, AI is now at a similar stage of development as space exploration was in 1958 when the UN formed its committee for the peaceful exploration of outer space, which led to the Outer Space Treaty (initially proposed by the US, UK and former Soviet Union in 1967, and since ratified by 107 countries). This treaty has been instrumental in providing the impetus and principles to underpin national guidelines and legislation in countries that have invested in developing their own space programs, covering a range of matters including "planetary protection" measures to prevent contamination of celestial bodies and Earth by foreign organisms.

Finally, there is simply a growing sense that the time has come for a more cohesive approach to AI oversight. Given the open research culture in the AI field, increasing availability of functional building blocks (e.g., machine learning models for image recognition, speech-to-text, translation; processing hardware), and the usefulness of AI to many applications, AI technology is spreading rapidly. If the world waits too long to establish international governance frameworks, we are likely to end up with a global patchwork that would slow the pace of AI development while also risking a race to the bottom. A self-regulatory or co-regulatory set of international governance norms that could be applied flexibly and adaptively would enable policy safeguards while preserving the space for continued beneficial innovation.

Key areas for clarification

This white paper highlights five specific areas where concrete, context-specific guidance from governments and civil society would help to advance the legal and ethical development of AI:

1. Explainability standards
2. Fairness appraisal
3. Safety considerations
4. Human-AI collaboration
5. Liability frameworks

While differing cultural sensitivities and priorities may lead to variation across regions, it should be feasible to agree on a high-level checklist of factors to consider. Longer term, working with standards bodies (such as ISO and IEEE), it may also be helpful to establish some global standards as 'due diligence' best practice processes in relation to developing and applying AI.

1. Explainability standards

Having an explanation for why an AI system behaves in a certain way can be a big help in boosting people's confidence and trust in the accuracy and appropriateness of its predictions. It is also important for ensuring there is accountability, not least in giving grounds for contesting the system's output. But delivering this in practice is not straightforward.

In thinking through what levels of explanation are acceptable, it is worth keeping in mind the standards applied to a human decision-maker in the same context. For instance, an oncologist may struggle to explain the intuition that leads him or her to believe they fear a patient's cancer has recurred. In contrast, an AI system in the same circumstance may be able to provide biomarker levels and historical scans from 100 similar patients as reference, even if it remains a struggle to fully grasp how the data are processed to predict an 80% chance of cancer.

As **Box 3** illustrates, there is no one-size-fits-all approach to what constitutes a reasonable explanation. The kind of explanation that is meaningful will vary by audience, since the factors emphasized and level of complexity that a layperson is interested in or can understand may be very different from that which is appropriate for an auditor or legal investigator. The nature of the use case should also impact the timing and manner in which an explanation can be delivered. Finally there are technical limits as to what is currently feasible for complex AI systems. With enough time and expertise, it is usually possible to get an indication of how complex systems function, but in practice doing so will seldom be economically viable at scale, and unreasonable requirements may inadvertently block the adoption of life-saving AI systems. A sensible compromise is needed that balances the benefits of using complex AI systems against the practical constraints that different standards of explainability would impose.

It might seem counter-intuitive, but giving lay users a detailed explanation may not necessarily be seen as helpful in practice. For instance, attempting to explain an AI system's prediction in terms of the underlying mathematical equations is unlikely to be decipherable by lay users, even if it were the most technically correct explanation. It is crucial to be guided by what people actually want and need. Sometimes their motivation may be more to have confidence that the system's inputs and output are fair and reasonable than to get a deep understanding of the calculation. And even if a thorough understanding of the model's functioning is sought, it can be overwhelming to receive in one burst. In our experience, shorter, higher-level explanations that people can probe for additional detail when and if they wish are often a more user-friendly approach.

BOX 3

What is a reasonable explanation?

There are many factors to take into account when thinking through the kind of explanation that is most appropriate in a given context. For instance:

Who is asking and what do they seek?
Different audiences will have vastly different needs. For example, lay users may want to know why an AI system made a specific decision relating to them, in order to have grounds to contest it if they feel that it is unfair or wrong. To be meaningful, this will need to be delivered to them in straightforward, non-technical language, which may limit the level of precision that can be provided. In contrast, expert staff at certification authorities will require a fuller, more technically detailed explanation of the system's functioning, so they can reassure themselves it meets expectations for reliability and accuracy at the general level. Similarly, there may be differences in the kind of explanation being sought. For example, an accident investigator will typically find a simple causal explanation most useful (e.g., the house was cold because the heating had been switched off), whereas a lay user might prefer an explanation that reflects the broader context (e.g., the house was cold because you selected the money saving option to turn off the heating when you're away).

When and where is it being delivered?
For instance, does the explanation involve sensitive or potentially embarrassing information, which the user might not want revealed in a public setting? Does the explanation need to be given upfront in real-time (which may present practical constraints), or is it sufficient to provide an explanation afterwards only on request?

What does it relate to?
The purpose of the AI system matters hugely. Systems being used to influence decisions of life-changing import, such as the choice of medical treatment, warrant much greater effort and depth of explanation than those performing tasks of minor consequence, such as making movie recommendations. The ease of contesting a decision and the availability of alternatives is another factor in determining how vital an explanation is. For instance, if an AI system were used by a restaurant to allocate tables, an explanation may not be important since unhappy diners could simply request a different table or go elsewhere. In contrast, an AI system used in parole hearings needs far greater explanation, because a person being held in custody has limited choice in the outcome and faces a higher hurdle to contest a decision.

How feasible is it to explain, technically and financially?
For some advanced AI systems there are limits (at the current state of research) on the extent to which it is possible to communicate in a human-understandable fashion how they function. Defining the target levels of accuracy and explanation required for a given application will help to identify which algorithms are appropriate to use (and not) in the design of that system. It is also important to recognize the cost dimension, in terms of the price and effort it takes to provide an explanation at scale, so that unreasonably detailed or rigid requirements do not block the adoption of valuable systems.

To aid in striking the right balance, we have been researching consumer satisfaction and understanding with different styles of explanations. **Box 4** illustrates our thinking so far on some of the hallmarks of a good explanation for lay users. The key seems to be to provide sufficient information to satisfy but not deluge; to impart a true sense of the system's complexity and the relative importance of different inputs without confusing or inadvertently misleading; and to give all this in a format that is comfortable for users to consume.

BOX 4

What are some hallmarks of a good explanation for users?

User-friendly explanations should be accurate, clear and specific, sensitive to context, and effective in improving overall understanding of the AI system. Key questions to ask:

Does the explanation accurately convey the key information underpinning the AI system's recommendation?

Obviously, an explanation that is incorrect or misleading is unhelpful. But determining accurate explanations for the output from complex AI systems can be tricky, since learning-based inferences are often made on the basis of multiple sources of information of varying influence. It can also be challenging to be specific enough such that a user clearly understands the inference source, especially when the inference is made based on the actions or attributes of other similar users as opposed to a user's own actions. In such cases if a variety of sources lead to an output, relying on the more influential sources can often yield a simpler, but still accurate, explanation.

Does the explanation boost understanding of the overall functioning of the AI system?

The more that users feel they understand the overall AI system, the more inclined and better equipped they will be to use it. Explanations can contribute to such insight, while ensuring that users don't inadvertently draw incorrect conclusions, such as confusing causation with correlation. Similarly, it may be helpful in some contexts to include an indication of how confident users should be in the accuracy of the AI system's output (e.g., users told that a result was only 70% likely to be correct would be more careful in acting on it than if told it was 98% likely). An important relevant design consideration is selecting where explanations will be placed during a user's interaction with a system, in order to offer the most transparency without overwhelming the user.

Is the explanation clear, specific, relatable and actionable?

To be helpful, an explanation needs to be understandable and provide sufficient information to give a sense of comfort, as well as provide grounds to appeal the outcome where appropriate. Ideally, explanations would refer to specific user actions logically indicative of the outcome, and allow users to grasp how their previous interactions led to a recommendation. A technically correct explanation which illuminates the mathematical model behind the decision but does not allow the individual to challenge the accuracy or fairness of its output in their case (e.g., in a parole risk assessment) may not count as a 'good' explanation. It is also important to recognize that an AI system may have multiple users, each with distinct roles and expectations (e.g., a healthcare application could be used by physicians, technicians, and patients all with varying expertise and stake in the results). Ideally, explanations should be tailored to the needs of different categories of users.

Does the explanation take appropriate account of sensitivities?

Some user-facing explanations can refer to sensitive information provided by or inferred about a user. Crafting accurate explanations that a user is comfortable with can be challenging, particularly for those who may be surprised at how much their aggregate data could reveal. It is also important to consider the setting in which the explanation is being provided. For semi-public settings (e.g., displayed on a shared screen like a TV, or in spoken format that may be at risk of being overheard) it may be better to surface only a general explanation, and follow up with a granular text explanation in a more private setting.

Assembling a collection of "best practice" explanations along with commentary on their praiseworthy characteristics (and conversely "poor practice" explanations with commentary on negative characteristics) would be a worthwhile collaborative exercise for policy stakeholders. This could include everything from effective user interfaces for delivering explanations, through to examples of documentation for experts and auditors (e.g., detailing performance characteristics, intended uses, and system limitations).

Obviously, it is not realistic to expect governments and civil society to provide guidelines on explanation standards specific to every instance in which AI systems may be deployed. However, doing so for some illustrative scenarios would provide industry with a calibration on how to balance the performance of various AI models that might be deployed within an AI system with the different standards of explainability required.

One way to begin could be to create a scale illustrating different levels of explanations. This scale could be used as a yardstick for setting minimum acceptable standards in different industry sectors and application contexts. For instance, if the potential bad effects of an error are small, then explanations are likely to be far less important than in cases where an error would be life-threatening[1]. Similarly, if users are able to easily avoid being subject to automated decision-making, there may be less expectation or need for an in-depth understanding.

When setting explanation standards it is vital to be pragmatic. Standards that are more difficult or costly to comply with could deter development of applications for which the financial returns are less certain. Requiring the most advanced possible explanation in all cases, irrespective of the actual need, would impose harmful costs on society by discouraging beneficial innovation. Appropriate standards of explanation should not exceed what is reasonably necessary and warranted. As an analogy, society does not expect an airline to explain to passengers why a plane is taking a particular algorithmically determined flight path — a similarly pragmatic and context-specific approach should apply to explanations for AI.

It is also important to factor in any potential tradeoffs versus system accuracy. For applications where "good enough" performance is sufficient, explainability might be prized over accuracy; in other instances where safety is paramount, accuracy might be prioritized so long as alternative mechanisms for providing accountability are in place. Alternatives could include the system being thoroughly tested by external auditors to identify various predictions in different situations; or offering contestability channels so that users can easily have their decisions re-assessed by a human, without prejudice, if they do not like the AI system's outcome. **Box 5** provides further suggestions on such alternative mechanisms.

BOX 5

Alternative ways to provide accountability

Sometimes, for technical or commercial reasons, it may not be feasible to provide an explanation of how an AI system functions that is sufficient to imbue confidence in its operation. In such cases, a combination of other methods should be deployed to test and monitor that the system is functioning properly. For instance:

Flagging facilities

Providing encouragement and making it easy for people to provide feedback when a system's output appears wrong or suboptimal is a crucial monitoring mechanism, and helps to pinpoint problem areas for deeper exploration. It is good practice no matter how much confidence there is in the system, because no system is ever perfect. It also empowers users to share their experiences and perceptions, making them feel heard and validated. This can engender more trust in the system and output decisions going forward if users feel their feedback is received and acted upon. Common techniques for flagging include user feedback channels (e.g., "click to report" button) and bug bounty programs where experts are incentivized to hunt for problems by getting paid (in terms of money and/or recognition) for each issue they report.

Avenues for contesting an outcome

If users have any doubts as to the accuracy of a system (and it is being used to do something that is significant), it is unlikely they will be willing to use it without a way to refute or appeal outcomes that they suspect are wrong. While having a channel for contesting results is helpful for all systems, it is particularly crucial for those in which no explanation for how they work has been given, as these are most likely to raise suspicions. The precise form that contestability mechanisms should take will vary by context and some will be more meaningful than others. For instance, being able to call and speak with a person who can provide a manual review and additional information is typically more robust than simply having an email address to send a complaint. But what is feasible will vary by context, and it will not always be possible (technically and financially) given the scale of likely requests to offer manual review.

Adversarial testing

Red team testing is a form of ethical hacking that involves assigning a team (which could be internal or independent) to do their best to find problems with the system. For example they could probe a system by inputting specific 'edge case' data to see if the output is as expected. The goal isn't only to find any areas where the system is broken, but also to stress test the surrounding processes including those related to reporting and contesting a decision.

Auditing

There are different kinds of audits that can be carried out, and these can be done by internal teams and (in some select instances) external bodies. For example, if there are legal standards of documentation to be met, auditors could review this paperwork to check compliance. In the case of an AI system such documentation might include details about the purpose of the AI system and its intended function and performance; information about the model architecture, datasets used in training and testing, internal checks made to ensure it was fit for purpose; and a review of organizational processes put in place to monitor system operations. There is also the potential for more investigative audits, where the system is interrogated by brute force — providing a range of inputs and reviewing the outputs to check they match the expected result. Related to these is the notion of auditing for disparate impacts (e.g., checking for disproportionately worse outcomes for marginalized groups). There are also a variety of techniques that may improve accountability even without access to the code base[2]. At the extreme, code reviews can be a possibility when there is sufficient expertise and doing so presents no risk to the security of the system, privacy of any underlying data, or in undermining intellectual property.

2. Fairness appraisal

Unfair stereotypes and negative associations embedded in algorithmic systems (deliberately or accidentally) can cause or amplify serious and lasting harm. These unfair biases can not only threaten social cohesion, but risk propagating unfairness in access to educational, economic or financial opportunities. Inadvertent differences in the quality or type of service provided to different groups can be just as damaging.

A complication, however, is that there are many conflicting definitions of fairness, whether decisions are made by humans or machines. For instance, is it fairer to offer an opportunity to any individual who satisfies the qualification criteria, or to an equal number of people from different population segments so as to avoid reinforcing historical disadvantage? Even for situations that seem simple, people can disagree about what is fair, and it may be unclear what optimal approach should dictate policy, especially in a global setting[3].

When building an AI tool to assist in decision-making it is necessary to make a choice upfront as to the precise fairness approach to adopt. Different technical approaches will result in models that are equitable in different ways. Deciding which to use requires ethical reasoning and is very context specific. Given the variety of perspectives and approaches to defining fairness, some definitions can directly conflict with one another, and others may promote equity only at the expense of accuracy or efficiency. However, if well implemented, an algorithmic approach can help to boost the consistency of decision-making, especially compared to the alternative of individuals judging according to their own internal (and thus likely varying) definitions of fairness.

This issue has particular resonance for policy makers, because algorithmic systems increasingly play a role in determining outcomes in public sector realms like the welfare or criminal justice systems. Governments can thus play a vital role in developing and modelling best practices, particularly in the articulation of frameworks to balance competing goals and definitions of fairness. For instance, it would be useful to have more clarity about the ways that the public sector makes fairness trade-offs in the context of specific decisions. While it is too early to expect to translate this into prescriptive metrics, it could still be useful guidance to others on how to grapple with similar issues.

More generally, governments and civil society could help by clarifying the relative prioritization of competing factors in some common hypothetical situations. For example, is it more fair to give loans at the same rate to two different groups, even if they have different rates of repayment, or is it more fair to give loans proportional to each group's repayment rates? At what level of granularity should groups be defined — and when is it fair to define a group at all versus factoring on individual differences? To what degree should algorithms consider or ignore individuals' gender, race, age, socio-economic circumstances, or other factors? While the answers will likely differ across cultures and geographies, having a shared understanding of the impact of such decisions, and some directional signposts, would be helpful for companies needing to make such tradeoffs.

Concerns about building fairness into algorithmic systems have helped to spark considerable research and policy efforts on fairness in AI. Google takes our responsibilities in this arena extremely seriously, not least in developing tools to tackle unfair bias, as highlighted in **Box 6**.

BOX 6

Google tools to help in tackling unfair bias

Google builds fairness and ethical considerations into the design, application and testing of our products. Our teams are leading the charge in creating tools that make it easier to surface bias, analyze data sets, and test and understand complex models in order to help make AI systems more fair. For example:

- **Facets:** Facets consists of two downloadable visualization tools to aid understanding and analysis of machine learning datasets[4]. Engineers can get a sense of the shape of their dataset using Facets Overview, and can explore individual observations using Facets Dive. The goal is to give engineers a clear view of the data they are using to train AI systems, helping to mitigate risk of bias. In 2018, Facets was used in the Gender Shades project of MIT Media Lab[5] to explore how well IBM, Microsoft and Face++ AI services guessed the gender of a face. By uncovering algorithmic bias across gender and race, the project has helped to motivate the development of inclusive and ethical AI.
- **What If Tool:** Building effective machine learning systems means asking a lot of questions. It is not enough to train a model and walk away. Instead, good practitioners act as detectives, probing to understand their model better. The What-If Tool is a TensorFlow plugin offering an interactive visual interface for exploring model results, without the need for writing any further code. For example, the What-If Tool lets model builders edit a datapoint to explore how the model's prediction changes, providing a sense of which factors are most influential in determining the result. It also supports exploration of different classification thresholds, taking into account constraints such as different numerical fairness criteria.
- **Model and Data Cards:** To reduce the risk of models developed for one purpose being applied in contexts for which they are ill-suited, we have developed a 'model card' documentation framework[6]. The ambition is for this documentation to accompany each released model and provide details of the model's intended purpose, how it performs in tests (e.g., for different genders, races, geographic locations, ages), and other relevant information. Similarly, to clearly delineate the makeup of a dataset, we propose outlining its unique characteristics, including where the data is from, the distribution of demographics represented in the dataset, and the source of labels (for labeled datasets)[7].
- **Training With Fairness Constraints:** Our researchers have developed state-of-the-art TensorFlow algorithms to train AI systems that satisfy standard desired statistical fairness goals, including demographic parity and equal odds. These advances are shared publicly with open-sourced TensorFlow software for anyone to use[8].

Rules set by policymakers also influence the extent to which fairness is able to be achieved and appraised. For example, inferring race can be essential to check that systems aren't racially biased, but some existing laws around discrimination and privacy can make this difficult. Similarly, while it might seem sensible to bar the inference of a person's gender to guard against unfair treatment, in practice doing so could inadvertently have the opposite effect, by making it harder to deliver reliable "mathematically fair" gender-neutral outputs. We urge policymakers and experts to work together to identify where this kind of inadvertent counter-intuitive harm arises, due to existing (or proposed) rules, and seek effective solutions.

Finally, it is important to also recognize and take advantage of the opportunities for AI systems to identify existing human and societal biases, and drive the world to become more fair. For instance, AI could be applied to analyze connections between input data and output predictions to surface any underlying biases that are embedded in existing processes. If these biases were determined to be unmerited, then decision-making practices could be tweaked in an effort to limit their effect. In the debate about the impact of algorithms on society and how they should be constrained, it is important to consider the potential for improving the consistency of decision-making and fairness of decisions.

3. Safety considerations

It is essential to take precautions against both accidental and deliberate misuse of AI with risks to safety. But this needs to be within reason, in proportion to the damage that could ensue and the viability of the preventative steps proposed, across technical, legal, economic, and cultural dimensions.

There are many challenges to the safety and security of AI systems. For example, it is hard to predict all possible AI system behaviors and downstream effects ahead of time, especially when applied to problems that are difficult for humans to solve. It is also hard to build systems that provide both the necessary restrictions for security, as well as the necessary flexibility to generate creative solutions or adapt to unusual inputs. **Box 7** illustrates some of the key areas of concern that need to be thought through when building an AI system.

However, it is important not to kid ourselves — no system will be perfect and problems will arise. The challenge is how to foster good safety practices and provide assurance that systems are sufficiently reliable and secure so that companies and society can feel confident in their use.

For ideas on how to do this we can look at analogies from elsewhere. For instance, researchers from the public, private, and academic sectors should work together to outline basic workflows and standards of documentation for specific application contexts which would be sufficient to show due diligence in carrying out safety checks (e.g., like for airline maintenance). See **Box 8** for an illustration of what we consider to be good practice in the safety testing and monitoring of automated industrial control systems, based on Google's experience deploying AI in our data centers.

There is also a need to take account of psychological factors. Sometimes there may simply be a need to (appropriately) foster user trust — for instance, the addition of a stop button and soothing voice recordings in early automated elevators provided crucial reassurance to those used to having elevator attendants.[9] The reciprocal concern is the risk of automation bias[10], in which regular users of a system can become complacent, and instinctively place more faith in its correctness than is warranted. Related is the lesser known inverse of algorithm aversion[11], which suggests that if a system is ever found to err, people lose confidence in it far more quickly than if a human had made the same error. This increases the risk of users choosing to ignore safety-critical guidance, even when a system is almost always correct, because of a single bad past experience.

BOX 7

Some key considerations to ensure AI safety

Safety problems related to accidents can be classified according to where in the process things went wrong. For instance[12]:

- **Is the objective function appropriate?**
  Many AI systems are models which seek to optimize a given objective. One problem to arise can be if limitations prevent measuring the desired objective in real time, or at all, so that a proxy metric needs to be used instead. How successful a model is at optimizing against the true objective will thus depend on how precise a match the proxy is in terms of its relationship to other variables. Other problems can occur if the chosen objective doesn't fully reflect the complexity of the environment, such that optimizing it has negative side effects or subverts the original intent. As an analogy, suppose a cleaning robot maker set the objective to remove visible dirt as fast as possible. If the optimal approach turned out to be hiding dirt under the carpet, or throwing away all visible dirty objects, this would be a failure in spirit even though it might satisfy the objective.
- **Has the exploration space been sufficiently constrained?**
  AI systems often come up with alternative better solutions because, unlike people, they are not constrained by ingrained assumptions about the typical way things are done. The flip side is that they typically lack common sense, and unless suitably constrained might inadvertently propose to try something that turns out to be harmful. For instance, a robot barista tasked with delivering coffee in the shortest time possible might (if given free rein) come up with the solution to throw the cup! This is one reason why simulations are often a sensible place to start when testing AI systems, so that they can be observed, and necessary restrictions put in place, to avoid such problems before use in a real-world setting.
- **Does the model's training reflect the current real world?**
  AI models learn from experience, based on the training data they are provided with initially, and (if permitted) the examples they encounter in use. Problems can arise if the training data is incomplete and misses some key aspects, or even if relevant aspects of the world have changed since the training data was collected. Part of due diligence to ensure the safety of an AI model is thus to pay close attention to the provenance and quality of the training data set, and adjust to mitigate against any shortfalls.

Safety problems caused by a lapse in security, or a clever hack, are more easily grouped based on the attack vector. Staying abreast of the latest research, solid development and design practices, and ongoing monitoring are the primary means of protection. For example:

- **Can the risk of data poisoning be mitigated?**
  AI systems that are continuously learning — rather than learning in lab conditions, and then having the underlying model frozen before real-world use — are likely to be at greatest risk of having the data they learn from corrupted. As a general rule, developers should think carefully about the data poisoning risks associated with having their AI systems learn in real-time in a real-world environment.
- **Has the AI system been adversarially tested?**
  This could be by a team of people playing at being adversaries, or an automated testing system in the form of adversarial learning — that is, using one network to generate adversarial examples that attempt to fool a system, coupled with a second network to try to detect the fraud. The more robustly a system is tested, the greater chance there is of finding points of weakness which can then either be fixed or (if that's not possible) monitored closely[13].

BOX 8

Overview of Google's approach to automating the control of data center cooling

We designed the AI system and underlying control infrastructure from the ground up with safety and reliability in mind, using eight different mechanisms[14]:

Continuous monitoring
to ensure that the AI system does not violate safety constraints.

Automatic failover
to a neutral state if the AI control system does violate the safety constraints.

Smoother transfer
during fallovers to prevent sudden changes to the system.

Two-layer verification
of the AI actions before implementation.

Constant communication
between the cloud-based AI and the physical infrastructure.

Uncertainty estimation
to ensure we only implement high confidence actions.

Rules and heuristics
as backup if we need to exit AI control mode.

Human override
is always available to override AI actions as necessary.

Extending beyond standard documentation, governments and industry could collaborate to establish safety certification marks that signify a service has been assessed as passing a set of checks that are relevant to particular uses (akin to CE certification on electrical products in Europe), for sectors which do not currently have such safety certification processes. For example, biometric recognition technology in smart lock systems could be tested against a representative, randomized dataset to ensure they exceed pre-set accuracy standards, before being certified safe for use. Or to reduce risks to safety from unexpected behavior, physical robots with baked-in AI could be required to have preset limits on how far their range of actions can veer from default settings without explicit user consent.

In setting benchmarks, it is important to factor in the opportunity cost of not using an AI solution when one is available; and to determine at what levels of relative safety performance AI solutions should be used to supplement or replace existing human ones. AI systems can make mistakes, but so do people, and in some contexts AI may be safer than alternatives without AI, even if it is not fail-proof.

In practice, the appropriate performance thresholds will also vary by context. If the damage from any errors is minimal, or in cases where it is difficult for people to complete the work within a set timescale, it may be deemed OK to use AI which falls below human levels of accuracy. In other situations, such a compromise may be ethically unacceptable, with AI required to show a significant jump in quality of output in order to justify its use.

The precise requirements to meet for safety certification in different scenarios would ideally be in line with internationally set standards, such as by ISO and IEEE. However, to be practicable, except in sector-specific instances (e.g., for medical devices), we would recommend safety certification to occur through a process of self-assessment similar to CE marks, backed up by existing sectoral governance bodies having the power to request documentation and carry out independent checks at a later date upon concerns.

Of course, regardless of any formal certification, ultimately it is companies and developers who are at the frontline of defense from bad actors. It is vital that they think carefully upfront about the kind of problems and attacks that their AI system is likely to face and their consequences, and continue to monitor the threat and update systems accordingly. This is the case regardless of the root cause — be it due to predictable system failures or unpredictable behaviors, unintentional misuse, or deliberate abuse and attack by bad actors. If the danger presented is severe enough, and there are not yet reliable ways to combat it, the right decision may be to simply not release the application until better protection mechanisms are available.

Finally, there is a broader and more philosophical question regarding safety in light of the multipurpose nature of AI. AI is a tool that can be applied with good or ill intent, and even well-meaning uses may turn out to be misguided in their real-world impact. There must be a balance between open publication and collaboration to accelerate access and progress, and thoughtful limitations and restrictions on openness to minimize harm.

This is a tradeoff with which Google has long grappled. For example, our decisions to open source Android as well as TensorFlow (Google's internally developed machine learning library) were made with careful deliberation, recognizing that the opportunities presented for beneficial use largely outweighed the potential impact of misuse by a small fraction of bad actors. In contrast, we have so far chosen not to offer a general API for facial recognition, due to outstanding technology and policy questions about this technology, and concerns that it could be used in ways that conflict with our AI Principles (e.g., to carry out extreme surveillance). Out of similar concerns we have also adjusted publication of some of Google's most cutting-edge AI work, either putting constraints on the models that are developed and shared, or even restricting the type of research we pursue.

As the ecosystem evolves, we continue to evaluate the tradeoffs between the benefits of openness and the risk of abuse for specific advances. While our preferred posture is to share in line with the open and collaborative nature of the AI research community, we do not do so naively. **Box 9** highlights key considerations we take into account when assessing whether and how to share our work. We welcome advice on how to best prioritize these conflicting elements, and where the AI research community should draw the line in sharing AI developments.

BOX 9

Ethical considerations in deciding whether to share Google AI advances

We generally seek to share Google research to contribute to growing the wider AI ecosystem. However we do not make it available without first reviewing the potential risks for abuse. Although each review is content-specific, key factors that we consider in making this judgment include:

- **Risk and scale of benefit vs downside –** What is the primary purpose and likely use of a technology and application, and how beneficial is this? Conversely, how adaptable is it to a harmful use, and how likely is it that there are bad actors with the skills and motivation to deploy it? Overall, what is the magnitude of potential impact likely to be?
- **Nature and uniqueness –** Is it a significant breakthrough or something that many people outside Google are also working on and close to achieving? Is sharing going to boost the capabilities of bad actors, or might it instead help to shift the playing field, so good actors are more able to offset the bad? What is the nature of Google's involvement — are we openly publishing a research paper that anyone can learn from, or are we directly developing a custom solution for a contentious third-party application?
- **Mitigation options –** Are there ways to detect and protect against bad actors deploying new techniques in bad ways? (If not, it might be necessary to hold back until a 'fix' has been found.) Would guidance on responsible use be likely to help, or more likely to alert bad actors?

4. Human-AI collaboration

"Human in the loop" is shorthand for systems which include people at one or more points in the decision-making process of an otherwise automated system. The challenge is in determining whether and where in the process people should play a role, and what precisely that role should entail, taking into account the purpose of the system and the wider context of its application (including, where relevant, a comparison to whatever existing process it is replacing). Ultimately, AI systems and humans have different strengths and weaknesses. Selecting the most prudent combination comes down to a holistic assessment of how best to ensure that an acceptable decision is made, given the circumstances.

However, making this determination is not straightforward. In some contexts, it is possible that a team of human and machine combined will perform better than either does alone. But in other situations it will be less clear-cut (e.g., a machine alone will perform many mathematical operations faster than in combination with a human), and an argument could be made that looping in a human would increase the risk of mistakes. Similarly, the degree of choice and control that users have has an impact on the ethics of fully automated processes. Delegating tasks and decisions to a machine is not bad, even in high stakes settings, so long as people have meaningful choice about doing so, and can revise their decision.

There are also considerations relating to fairness. While a lot of attention has focused on the risk that poorly designed and applied AI systems might have baked-in unfair bias, the same risks are true for people. This is not to imply that there is no problem with biased AI; but rather to point out that there may be instances where a person is likely to be more biased than an AI system. In such cases, well-designed, thoroughly vetted AI systems may reduce bias compared with traditional human decision-makers.

In addition, there are factors beyond system accuracy, speed, and scale to consider. For instance, some have argued that allowing certain kinds of life-determining medical decisions to be made solely by machines may fail to respect the right to human dignity. Similarly, if empowering or educating people is a high-priority operational goal, this may have implications for the nature of the role that people are assigned in the AI collaboration process.

Looking holistically, people are central to an AI system's development and likely to remain so. From the beginning stages of problem and goal articulation, through to data collection and curation, and model and product design, people are the engine for the system's creation. Even with advanced AI systems able to design learning architectures or generate new ideas, the choice of which to pursue should still be overseen by human collaborators, not least to ensure choices fall within an organization's legal and financial constraints. Similarly, people play a vital role in the upfront verification and monitoring of a system, such as choosing which tests to run, reviewing results, and deciding if the model satisfies the performance criteria so as to enter (or remain) in real-world use. And of course, human users provide essential feedback to improve AI systems over time.

While it would be hubris to presume to know the optimal human-AI collaboration structure for every situation, there are early inklings enough to make a start at outlining some guidelines (see **Box 10**).

BOX 10

Considerations for successful human-AI collaboration

Design for the different strengths of people and machines – Machines have many great qualities — they never forget (unless they are designed to), and they can crunch numbers and scan documents faster than a person without getting bored or impatient. But in comparison to people, machines are less capable of picking up on emotional nuances; they lack common sense; and they need more detailed instruction and hand-holding with new tasks. More fundamentally, machines will never be able to bring a genuine humanity to their interactions, no matter how good they get at faking it. Such differences should be front of mind when thinking about the kind of tasks and settings in which to deploy an AI system to amplify and augment human capabilities.

Successful collaborations are built on communication – The more people know about others' mindsets and the reasoning behind their actions, the more likely it is they will work effectively together, because trust is built. The same is true of people and machines. That is why it is so helpful for AI systems to be able to explain their reasoning and the key factors that led to a certain output. For AI systems which are probabilistic by nature, this should also include an indicator for how much the output should be trusted (e.g., a confidence score for predictions of a medical diagnosis). Just as important as having an explanation is the way that it is delivered. If an explanation is unclear or too hard to find it loses potency, since if explanations are seldom reviewed then a vital opportunity to "sanity check" is lost. Similarly, the greater the scope for people to provide relevant context, the more nuanced and accurate the machine's output can be.

Flexibility in role assignment is a boon –Just as with teams of people, it is helpful for there to be fluidity in the nature of the roles played by a person working alongside a machine, especially in safety-critical situations. This ensures that people retain the skills and confidence to carry out tasks, making them psychologically more willing to question a machine's output if they feel there is something wrong. If certain tasks are permanently delegated to machines, people using them will naturally adopt a more laissez faire approach, making it less likely they will spot errors. This may be fine in settings where the risks of malfunction are low and other safety mechanisms are in place, but less so for high-stakes use cases, especially if these involve complex procedures with a lot of variability. More generally, careful thought should be given to when and how issues should be escalated to a person to respond.

Design processes with human psychology in mind – People's natural emotional and cognitive tendencies should not be overlooked when deciding the roles and settings for AI systems. For example, if an alarm system is set so that there are a lot of false positives, over time it is likely at best to be seen as an irritant, and at worst to be assumed to be wrong (even when it is not). Similarly, if management introduces an AI system without consultation, it may fuel resentment among those who need to work with the system, making them less engaged and maybe even inclined to seek ways to sabotage its effectiveness. Expertise from the fields of user experience design (UX) and human-computer interaction (HCI) can shape the design of more effective and satisfying models of AI-human cooperation in the workplace.

Governments may wish to identify red-line areas where human involvement is deemed imperative. For instance, for ethical reasons we would suggest that people should always be meaningfully involved in making legal judgments of criminality, or in making certain life-altering decisions about medical treatment. It would also be useful to have broad guidance as to what human involvement should look like — for example, an evaluation of common approaches to enabling human input and control, with commentary on which are acceptable or optimal, supplemented by hypothetical examples from different contexts. See **Box 11** for some initial thoughts on how this might be calibrated.

BOX 11

Ilustration of variance in AI system operator roles

There can be great variation in the nature of an AI system operator's role. Three key factors include the level of awareness the operator has, what scope they have to provide input, and their level of control. This table lays out some initial thoughts on how this might be calibrated:

	Level of awareness	Level of input	Level of control
0	Little knowledge of how the AI system works, beyond its existence and overall purpose, and ability to observe inputs and outputs	No facility to alter or provide additional input other than via upstream processes (e.g., information provided by user, or from historical records)	None
1	General understanding of the way the AI system operates and criteria for its successful and safe functioning. Awareness of the most common factors that can cause mistakes	Facility to tweak initial inputs to the AI system, and provision of guidance on when doing so might be appropriate	Upfront choice over whether to engage the AI system
2	Detailed understanding of the AI system's operation and criteria for its effective operation. Expert training and tools provided to monitor and check for potential problems that may arise	Same as 1 plus more detailed information about which factors are the key influences over the AI system's outcome and their relative sensitivities (e.g., in the form of counterfactuals[15])	Same as 1 plus ability to intervene and choose not to use the AI system's output
3	Same as 2 plus the addition of forensic auditing facilities enabling investigation of specific instances (rather than only overall model operation)	Same as 2 plus the ability to alter the weightings that describe the relative importance of different factors	Same as 2 plus the ability to prescribe custom operational boundaries (e.g., if someone has been a loyal shopper at your store for 5+ years and is in good standing, to never deny their request to return low value purchases)

More generally, guidance would be useful about the extent to which people should be able to switch off an AI system to which they have previously chosen to delegate a task. In the case of consumer-facing services, we believe there should always be an ability to avoid engaging with an AI system (even if in practice this means missing out on the benefits enabled by the service, or requires not participating in certain activities). However, with regards to enterprise AI systems it is more complex, since switching it off could have legal consequences or inflict harm on others. For instance, switching off an AI monitoring system without putting in place a credible alternative could undermine public safety if so doing increased the risk of accidents. In such cases, we would propose there needs to be upfront consideration of backup options, and a clear approval process prior to a switch off being initiated, including consultation with affected parties.

Safety concerns are one of the main reasons people give for seeking to ensure there is a "human in the loop" in AI implementations. This is based on the perception that having a person overseeing an AI system's recommendations will provide a fail-safe mechanism to protect against mistakes. Unfortunately in many instances this is a fallacy. In practice, it is seldom scalable to have a person checking every recommendation from an AI's system, so oversight ends up being limited to just those that the system is less sure about (i.e., that fall below a probability threshold). Thus fundamental mistakes about which the AI system is confident will be missed.

Process designers must also contend with the realities of human psychology. On one hand, there is the risk that people may misjudge and overtrust in their own capabilities. On the other, there is the risk that people who have spent a long while working with a system where errors are rare (as should be the case for production AI systems) become naturally less inclined over time to question the system's accuracy due to automation bias — aka the "computer says yes" syndrome[16]. This is made worse when reviewers are under pressure and there is a cost to reporting a potential problem, be it the time taken to file a report, or the damage to their reputation for flagging something that turned out after examination to be a false alarm. There are ways to reduce such risk (e.g., not telling the reviewer what the system recommended until they have come to their own conclusion; setting reviewers a quota they must meet for queries; providing bounty rewards for finding errors), but it requires careful planning of processes and organizational structure to implement. **Box 12** summarizes how we approach this challenge at YouTube.

BOX 12

YouTube case study for human-AI collaboration

At YouTube, we work hard to maintain a safe and vibrant community. Our Community Guidelines set the rules for what content we don't allow on YouTube.

We have long used a mix of technology and humans to deal with harmful content. Our technology notifies us of content that may violate our policies; our community of users also flags content to us for review using the various reporting options available on the platform. Content flagged by technology and users is reviewed by teams based in multiple locations around the world so we can take appropriate action in a timely manner. We also use technology to prevent exact reuploads of content that we have determined to be in violation of our policies.

As AI technology has advanced, it has become a powerful tool to help detect this content quickly at scale for some of the most harmful varieties, like violent extremism and child exploitation. At the same time, AI-based systems still make many errors in context-sensitive tasks, which is why we strive to keep a human in the loop when evaluating new material. This human element preserves accountability while also identifying classifier error and developing better training data, improving the model for future iterations.

Between July and September 2018, 81% of the 7.8 million videos removed from YouTube were initially flagged by our AI systems. Of removed videos first flagged in this way, 74% had no views. Well over 90% of videos uploaded in September 2018 which were removed for child safety or violent extremism violations had been viewed fewer than 10 times.

Regardless, it is likely there will always be sensitive contexts where society will want a human to make the final decision, no matter how accurate an AI system is or the time/cost benefits of full automation. **Box 13** provides examples of some possible factors to consider. We urge regulators to work with civil society and other stakeholders to agree on the characteristics of such instances on a sector-specific basis.

BOX 13

Factors to consider relating to sensitive AI use cases

While every case needs to be evaluated on its merits, some categories of issues will require detailed protocols:

Does the decision materially affect someone's life? – AI systems being used to determine credit, access to housing or education, choice of medical treatment, decisions of criminality, and similar high-stakes decisions may have a substantive and irrevocable negative impact on those affected. Fully delegating such decisions to machines —or even giving the perception that is what is happening (regardless of truth) — may fairly be seen as an affront to human dignity. However, a pragmatic balance is needed, since requiring every decision in these areas to be made manually would be inefficient, and untenable to serve people in a timely manner at scale.

Does the decision impact a new versus a pre-existing benefit? – Where feasible, it is advisable to trigger a human review prior to any action being taken if an AI system were to recommend reducing the level of service provided to an existing customer.

To what extent can a decision be contested? – In practice, people tend to be far less concerned about the process used to reach a decision if there is an option of meaningful human review.

Does it involve a situation that could impinge on the underpinnings of society or human rights, in a local context? – For example, in close elections when a recount is required, standard practice is often for that to be done by hand, not machine. Similarly, in criminal trials, a final decision of guilt or innocence, and the form of punishment, should never be delegated to an AI system — even if shown to have the potential to reduce bias. More generally, there is a worry that AI systems might inadvertently foster cognitively harmful habits in some people (e.g., extreme compulsive use of social media that disrupts sleep and mental health), or to undermine humane societal norms (e.g., if machines replaced physician-patient interactions, rather than assisting and supplementing them).

5. Liability frameworks

Organizations should remain responsible for the decisions they make and the manner in which they act on them (whether using AI or humans or both). For the reasons laid out in **Box 14,** it is not appropriate for moral or legal responsibility to be shifted to a machine. No matter how complex the AI system, it must be persons or organizations who are ultimately responsible for the actions of AI systems within their design or control.

Things are less clear-cut, however, in regard to expectations of behavior that apply to AI providers. Few organizations outside of the tech arena will develop their AI systems solely using in-house expertise. Most commonly they will collaborate with third-party AI providers, who have the expertise and tools to help design and operationalize an AI system that meets the organization's needs, far faster and with higher quality. The onus is on AI providers to help their clients to understand the risks inherent in using AI systems, so they can make educated decisions on how to mitigate and monitor for them (e.g., warning about the performance limitations of off-the-shelf models). Naturally, however, different contributors to any complex enterprise system may not have full visibility into all applications.

Governments may wish to work with other stakeholders to provide greater clarity on the expected behavior of providers of AI services, and of clients using AI for applications in specific fields. For example, should there be additional precautions for certain categories of end-use and sector? If evidence of misuse emerges, how should AI providers respond if clients are not willing to address the concern? Of course, any such requirements would need to be backed by new norms, standards, regulations, or laws in order to be consistently applied and useful to all providers and clients.

BOX 14

Why legal personhood for AI is a bad idea

Many of the calls for legal personhood for robots or AI are based on a superficial understanding and overvaluation of the actual capabilities of and objectives for even the most advanced AI systems. In April 2018 a group of leading AI experts and roboticists convincingly laid out their views on why this was a bad idea in an open letter to the European Commission[17]. Google shares this opinion for the following reasons:

- **It is unnecessary:** There will always be a natural person or corporation liable within existing laws and legal frameworks. Legal personhood is a solution to a problem that does not exist.
- **It is impractical:** Even if it was possible to come up with a workable definition of robots or AI that warrant legal personhood (which is far from a given), it would be impossible to hold such entities accountable for violations of their obligations. To put it another way, how can a machine that lacks consciousness or feelings be punished?
- **It is immoral:** Responsibility is an intrinsically human property. It is morally inappropriate to shift responsibility to "synthetic persons" in the form of machines or code.
- **It is open to abuse:** It would make it easier for bad actors to shield themselves from liability for illegal activities performed by machines they had created.

More generally, there has been some debate about whether the emergence of AI requires the creation of new laws regarding liability. Countries already have long-established legal frameworks that provide guidance in this arena — not least contract, tort, consumer protection and criminal law — although which frameworks come into play, and to what degree, may vary across sectors and use contexts. Seeking redress within complex value chains, such as the car manufacturing industry, has been commonplace for many years, and existing laws regarding liability seem largely fit to also deal with AI technologies.

However, while in many cases this tried-and-tested approach to liability will work, there may be times when it fails. Untangling the causal strands of who was responsible for what can be tricky even in human-only situations; this can become far more difficult as complex algorithms with various human touchpoints are added. There is a growing concern over how best to ensure that end users of complex AI systems are adequately protected if there are so many contributing factors to what happened (including potentially even the autonomous actions of a machine) that responsibility becomes diffuse, and it is hard to reliably assign blame for problems.

For example, the European Commission is currently evaluating the existing liability framework for its fitness in the light of so-called "emerging digital technologies" that include AI systems[18]. One approach being evaluated is the extension of the scope of "products" to include stand-alone software as well as services, which would make AI systems subject to strict liability[19]. Some have even gone so far as to moot an extension of the concept of a defective product to include the provision of "defective information". Other approaches involve a joint[20] (strict) liability of all actors within the network, or the reversal of the burden of proof as far as an element of negligence is still required.

While such approaches might indeed strengthen the legal position of the end users of AI systems, they also come with considerable downsides. Strict liability would bring increased exposure to legal uncertainty, as it would mean that anyone involved in making an AI system could be held liable for problems they had no awareness of or influence over. It could lead to misplaced responsibility, if the AI system was not actually at fault and just a conduit, rather than the original source of harm. Burdening AI system manufacturers with such a risk would additionally have a chilling effect on innovation and competition. Similarly, a blanket approach to holding systems liable for "defective information" would also risk curtailing the expression of ideas (akin to holding an app providing driving directions liable for not having known that a road was flooded).

Joint liability is also problematic because it could reduce incentives for smaller players in the value chain to behave responsibly, since they would be less likely to be targeted if something went wrong, as plaintiffs would seek compensation from bigger players. Introducing joint liability thus could have the perverse impact of reducing the overall safety of AI systems.

Overall, Google recommends a cautious approach for governments with respect to liability in AI systems, since the wrong frameworks might place unfair blame, stifle innovation, or even reduce safety. Any changes to the general liability framework should come only after thorough research establishing the failure of the existing contract, tort, and other laws.

Should a need for action be identified in areas that involve increased risks for end users (e.g., healthcare and health research, financial services, road traffic, aviation) this should be addressed in a sector-specific manner, with new regulation added only where there is a clear gap and in a way that minimizes overspill. Sector-specific safe harbor frameworks or liability caps (as with medical malpractice, orphan drugs, or nuclear energy plants) are also worth considering in domains where there is a worry that liability laws may otherwise discourage societally beneficial innovation.

For example, suppose in Europe it was deemed desirable to have a strict liability framework for AI systems used to determine medical treatment. The simplest way to achieve this would be to update European medical device regulation. Doing so would not alter the legal standing of physical medical devices (which already face strict liability under the Product Liability Directive), and there is already precedent to indicate that software can be considered as a medical device in Europe. Updating such sector-specific regulation, rather than making sweeping changes to general product liability frameworks, would allow for more precise targeting of changes. Where relevant, safe harbor provisions could encourage innovation needed to advance the state of the art in tackling high-priority diseases.

Overall, there are a variety of possible liability regimes that could be applied to AI systems. Each has pros and cons, fuelling lively debate in legal and policy circles. As technology evolves, so too should law — but making changes to such a fundamental business and societal underpinning as liability should be done thoughtfully and conservatively, in response only to evidence of a clear gap. And no matter what liability regime is in place, it is vital to ensure there are means of exoneration for actors providing evidence that they did not proximately cause a reasonably foreseeable harmful outcome.

Alternatively, in some circumstances (e.g., where the costs of adjudicating liability are high, and the deterrence value of individualized liability is low), governments and insurers may want to consider compulsory insurance programs. Google would support discussions with leading insurers and other stakeholders on appropriate legislative models.

In Closing

This paper highlights what Google considers to be some of the critical current questions in the debate on AI governance. We hope it is useful as a practical contribution to the lively debates on AI oversight now underway in many forums around the globe.

Overall, Google believes the optimal governance regime is one that is flexible and able to keep pace with developments, while respecting cultural differences. We believe that self- and co-regulatory approaches will remain the most effective practical way to address and prevent AI related problems in the vast majority of instances, within the boundaries already set by sector-specific regulation.

However, we recognize that there are some instances where additional rules would be of benefit, and we look forward to engaging with governments, industry practitioners, and civil society on these topics. Some contentious uses of AI could have such a transformational effect on society that relying on companies alone to set standards is inappropriate — not because companies can't be trusted to be impartial and responsible, but because to delegate such decisions to companies would be undemocratic.

These contentious uses share two commonalities. First, they represent a major and irrevocable shift in the scale of possible harm that could be inflicted. This could involve anything from a new kind of weapon to an application that fundamentally overhauls everyday norms (e.g., the ability to be anonymous in a crowd, or to trust in what you see). Second, there is much debate over where the lines should be drawn in terms of what is permissible, and by whom, with reasonable arguments on both sides. For instance, how should societies trade off the opportunities for AI-powered surveillance to reduce crime or find missing persons, with the implications it will have for privacy and human rights?

While this white paper is focused on pressing questions regarding the implementation of AI generally, we recognize questions on contentious use cases are important, and plan to share our developing perspectives on such uses in the near future. Ultimately, while experts can advise on technical and practical constraints, and can even decide not to pursue certain legal applications, the decision on how societies should employ such uses (or not) rests with government.

On a related note, there are a myriad of national and regional initiatives underway seeking to establish organizational structures for AI oversight. We support the collaborative and consultative process that many are pursuing, and encourage stakeholders everywhere to participate. As initiatives progress, we hope to find opportunities for Google to continue to listen to, learn from, and contribute more actively to the wider discussion about AI's impact on society.

End notes

1 For more on the scope and scale of possible harms, see Future of Privacy Forum's 2017 report on "Unfairness By Algorithm: Distilling the Harms of Automated Decision-Making". Available online at https://bit.ly/2C2G20w

2 Kroll et al., "Accountable Algorithms," 2017. Available online at https://bit.ly/2SzpggH

3 For more examples see the video tutorial on "21 fairness definitions and their politics" by Arvind Narayanan. Viewable at https://youtu.be/jIXIuYdnyyk

4 More on the Facets tools can be found at https://pair-code.github.io/facets/

5 More information about the Gender Shades project is provided at https://bit.ly/2OWTcWi

6 For more see "Model cards for model reporting" by Mitchell M et al (2018). Available online at https://arxiv.org/abs/1810.03993

7 For more see "Datasheets for datasets by Gebru T et al (2018). Available online at https://arxiv.org/abs/1803.09010. An example of this in action is the data card produced for the Open Images Extended dataset, viewable at https://bit.ly/2sh4czU

8 The research paper outlining Google's work on training with fairness constraints is available online at https://arxiv.org/pdf/1809.04198.pdf

9 "Remembering When Driverless Elevators Drew Skepticism" featured on NPR in July 2015 and available online at https://n.pr/2pXHig0

10 "Complacency and Bias in Human Use of Automation: An Attentional Integration", by Raja P et al 2010. Available online at https://bit.ly/2AdwNLg

11 "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err", by Dietvorst B et al 2014. Available online at https://bit.ly/2ErYeFo

12 A fuller and more technically oriented overview of safety considerations is in the article "Building safe artificial intelligence: specification, robustness, and assurance" by Ortega P et al (2018). Viewable online at https://bit.ly/2R5bgLb

13 For more background see "Adversarial perturbations of deep neural networks" by Warde-Farley D et al 2016 and "Adversarial examples: Attacks and defenses for deep learning" by Yuan X et al 2017. Viewable online at https://bit.ly/2RePNmN and https://arxiv.org/abs/1712.07107 respectively.

14 For more information see blogpost on "Safety-first AI for autonomous data center cooling and industrial control" at https://bit.ly/2AzCT8O

15 Counterfactuals are the most similar point where the system would predict a different result. A statement such as "if your weekly income had been $2500 instead of $2400 you would have been granted the loan" is an example of something that could be a counterfactual. For more on this topic see Wachter S et al. 2018 "Counterfactual Explanations without Opening the Black Box: Automated decisions and the GDPR" at https://arxiv.org/abs/1711.00399

16 "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust" by Hoff KA et al 2015 provides more detail on factors which influence people's trust in automation. Available online at https://bit.ly/2SX48Az

17 April 2018 letter from AI experts objecting to the notion of legal personhood for AI and robots is viewable online at https://bit.ly/2xfMToe. For more on this topic see Bryson, J.J. et al. 2017 "Of, for, and by the people: the legal lacuna of synthetic persons" at https://bit.ly/2Auvjft

18 The European Commission Staff Working Document on liability for emerging digital technologies was published in April 2018 and is viewable online at https://bit.ly/2I1zIMH

19 Under a "strict liability" regime, the injured person does not have to prove a fault of the defendant. Under the EU's Product Liability Directive 85/374/EEC the injured person, however, carries the burden of proof of the defect in the product, the actual damage and the causal link between the defect and the damage.

20 "Joint liability" allows several defendants to be sued for one tort they have caused. The plaintiff is allowed to collect the full amount of damages from any single defendant regardless of relative fault of each defendant.

Google