Google

Recommendations for Regulating AI

Background

Google has long championed AI. Our research teams are at the forefront of AI development, and we've seen firsthand how AI can enable massive increases in performance and functionality. AI has the potential to deliver great benefits for economies and society — from improving energy efficiency and more accurately detecting disease, to increasing the productivity of businesses of all sizes. Harnessed appropriately, AI can also support fairer, safer and more inclusive and informed decision-making. We are keen to ensure that everyone and every business can benefit from the opportunities that AI creates.

AI will have a significant impact on society for many years to come. That's why we established our AI Principles (including applications we will not pursue)[1] to guide Google teams on the responsible development and use of AI. These are backed by the operational processes and structures necessary to ensure they are not just words but concrete standards that actively impact our research, products and business decisions to ensure trustworthy and effective AI application.

But while self-regulation is vital, it is not enough. Balanced, fact-based guidance from governments, academia and civil society is also needed to establish boundaries, including in the form of regulation. As our CEO Sundar Pichai has noted, AI is too important not to regulate. The challenge is to do so in a way that is proportionately tailored to mitigate risks and promote reliable, robust and trustworthy AI applications, while still enabling innovation and the promise of AI for societal benefit. Since the publication of Google's whitepaper on AI governance[2] we have provided input to multiple government consultations[3] and engaged in many discussions about the opportunities and challenges of regulating AI. This paper aggregates our foundational principles for regulating AI, as well as providing detailed commentary on key topics that have become a focus for attention. See **Box 1** for a topline overview.

Box 1

Topline recommendations for regulating AI

General approach:

1. Take a sectoral approach that builds on existing regulation
2. Adopt a proportionate, risk-based framework
3. Promote an interoperable approach to AI standards and governance
4. Ensure parity in expectations between non-AI and AI systems
5. Recognise that transparency is a means to an end

Implementation practicalities:

6. Clarify expectations for conducting risk assessments
7. Take a pragmatic approach to setting disclosure standards
8. Workable standards for explainability and reproducibility require compromise
9. Ex-ante auditing should centre on processes
10. Ensure fairness benchmarks are pragmatic and reflect the wider context
11. Prioritise robustness but tailor expectations to the context
12. Be wary of over-reliance on human oversight as a solution to AI issues

General approach

There can be no sensible "one size fits all" approach to regulating AI because there is no single use case for AI. AI is a multi-purpose technology which takes many forms and fulfills many purposes, spanning a wide range of risk profiles. Notably, it may be some of the AI capabilities and applications considered highest risk that are also of highest value to society. It's important therefore to take a holistic approach in tailoring regulation to reflect the role being played by AI within a given operational and sectoral context, as well as the nature and probability of possible harm. While it is obvious that some uses of AI will warrant extra scrutiny and safeguards, in parallel there should be clear acknowledgment of the opportunity costs of not developing and using AI.

AI may present some novel challenges for regulators, including building up new expertise, but they do not need to reinvent the wheel in every respect. There are general foundational principles of good regulatory design that should be part of any approach to regulate AI, which are outlined in this section.

1. Take a sectoral approach that builds on existing regulation

Just as there is no horizontal regulation across the fields of mathematics or biology, the focus in AI regulation should be on specific applications of AI — not the science of AI itself. There is an immense diversity of AI applications across almost all sectors of society—healthcare, financial services, transportation, energy, science, retail, agriculture, logistics, manufacturing, and beyond—and their impact on people and organizations are not the same. Additionally, many "AI issues" are actually issues common to the operation of any complex software already used by retailers, banks, insurance companies, manufacturers, and others. Consequently, AI regulation is likely best addressed first through sectoral approaches that leverage existing regulatory expertise in specific domains, rather than one-size-fits-all horizontal approaches.

Governments should therefore look first to existing regulatory experts, frameworks, and instruments that may encompass AI applications. Such sectoral experts typically will be well-positioned to assess context-specific uses and impacts of AI and to determine whether and how best to regulate them, although in some circumstances additional resources may be required, including internal technical AI expert capacity. For instance, health-focused agencies are best positioned to evaluate the use of AI in medical devices. Similarly, energy regulators have expertise in evaluating the use of AI in energy production and distribution. Having consistency in oversight and the expectations for human and machine actors performing the same task will also help to reduce the risk of artificial protectionist constraints being imposed, unless there are justifiable grounds for difference.

If action is determined to be necessary, to avoid duplication and speed implementation Google recommends expanding established due diligence and regulatory review

processes to include the assessment of AI applications. If there are instances where the AI application in question is not obviously covered by existing regulations, clear guidance should be provided as to the "due diligence" criteria companies should use in their development processes. This would enable robust upfront self-assessment and documentation of any risks and mitigation strategies, and could also include further scrutiny after launch.

2. Adopt a proportionate, risk-based framework

AI is not risk-free, but when developed and used responsibly it can help reduce a vast array of risks inherent in everyday life. Conceptually, Google supports a risk-based approach to any new regulatory framework, but it is vital to ensure that it is targeted at the right use cases, taking into account the likelihood of harm and not just the severity of the harm, as well as consideration of the cost of not using AI in terms of forgone benefits.

A proportionate approach is necessary, balancing potential harms with the many social and economic benefits promised by AI, and clearly acknowledging the opportunity costs of not using AI in a specific situation or developing AI with particular capabilities. It's important to acknowledge that there are flaws in existing (non AI) approaches, and if an imperfect AI system were shown to perform better than the status quo at a crucial life-saving task, it may be irresponsible to not use the AI system. In instances where the alternative of not using AI poses greater risk than the risk posed by deploying an AI system, the regulatory framework should not discourage AI's net beneficial use.

It's also important to reflect the wider operational context when assessing the level of risk. Organizations using AI will have more encouragement to invest in additional mitigations and safeguards to reduce risks if doing so reduces the regulatory burden.

3. Promote an interoperable approach to AI standards and governance

Given the cross-border nature of the digital economy, AI regulatory frameworks and technical standards should ideally operate across nations and regions. Increased global alignment on AI regulation, including in the context of trade, will help to facilitate the adoption, use, and interoperability of AI technologies across different jurisdictions.

Internationally recognized voluntary consensus standards can serve as the basis for robust self- and co-regulatory regimes, as guideposts for regulators, and even as the regulatory standards themselves if incorporated by reference. Because such standards are based on a broad and deep foundation of expertise from a wide variety of industry and civil society perspectives, they can be flexible and nimble in a way that static regulation cannot, evolving over time as the technologies innovate and change.

There are a variety of efforts underway to establish internationally recognized standards for AI, including within ISO and IEEE, as well as industry-driven initiatives such as MLPerf.

While these efforts can provide helpful pointers and highlight key areas for attention, as they progress it is important they are able to evolve in line with the still-rapid developments in underlying AI technologies. Ultimately it is unlikely that a single set of standards will emerge to suit all circumstances, but rather that there will be multiple families of standards. Accordingly, as in similar domains such as cybersecurity, regulators should avoid the temptation to "pick winners" and instead allow flexibility for the optimal standards approach to be chosen for each specific context.

More generally, Google encourages policymakers to pay attention to the work of the OECD and the Global Partnership on AI (GPAI), two fora that are emerging as international clearinghouses for progress on AI governance. By taking coordinated (even if not identical) approaches, regulators can avoid adopting measures that inhibit cross-border research or disproportionately impact AI applications created in other countries. Cross-border cooperation among regulators is also critical to helping governments jointly develop and deploy AI to address global challenges related to public health, humanitarian assistance, sustainability, and disaster response.

4. Ensure parity in expectations between non-AI and AI systems

Like any system, including human-based processes, AI systems are not perfect. They do, however, offer the opportunity to dramatically improve on current human-based decision making. Thus, the operational benchmark for AI systems should not be perfection, but instead the performance of comparable current processes (if existing) or an available human-powered alternative.

There is a real risk that innovative uses of AI could be precluded by demanding that AI systems meet a standard that far exceeds that required of non-AI approaches. Sometimes this may be deliberate due to artificial protectionist constraints, but more often it is likely to be due to a lack of understanding about hidden flaws in existing non-AI decisions, and people's natural tendency to be more forgiving of mistakes made by a human vs a machine. To help offset this, Google would recommend that there should be parity in terms of expectations between AI and non-AI approaches, unless there is a clear justification put forward as to why it should differ for a particular use case and context.

A key area in which this principle should apply is setting minimum performance standards. A sensible starting point would be to expect AI systems to match or exceed similar accuracy and fairness standards as current approaches. There may however be good reasons to deviate from this. In some situations a lower level of accuracy may be acceptable — such as if an urgent response is needed, the cost of inaction is high, and there are simply not enough qualified people on hand to do the job (e.g., helping triage medical screening in crisis settings). In other contexts the reverse may hold — such as if there are plenty of qualified people happy to do the work, using an AI system might only be justified if it was shown to perform significantly better (e.g., self driving cars that have far fewer accidents than human drivers). Similarly, fairness is a vital consideration — even

if an AI system performs more accurately and reliably across the general population, that may not be sufficient to justify its use if it performs significantly worse than existing approaches for certain subgroups.

A similar principle of comparison can also be applied to expectations of transparency and explainability for AI systems. However in so doing, it is important not to exaggerate the standards met by human-powered systems. There are many settings where an explanation is not required of human decision-makers, and even if an explanation is provided, there is no means of ensuring that it accurately represents the key factors influencing a person's decision (e.g., people may opt to withhold mention of certain factors, or there may be unconscious bias).

5. Recognise that transparency is a means to an end

Transparency is not an end in and of itself — it is a means by which to enable accountability, empower users, and build trust and confidence. In designing transparency requirements, governments should consider what they are trying to achieve and how best to meet those goals in a given context.

For example, different stakeholders will require different forms of transparency under different circumstances. Requirements should be tailored to ensure that the information is actionable and presented when stakeholders want it, in terms they can understand, and without extraneous details that can be distracting or confusing. In some contexts, detailed information about individual decisions may be important, whereas in others general information about how systems work or how they are developed may be more beneficial or appropriate.

Regulators should also carefully balance the desire for transparency with other important equities, for example speed, safety, security and privacy. Importantly, some forms of transparency that are intuitively appealing can carry some of the most significant risks, and provide little actual benefit in terms of enabling accountability and building trust. For example, disclosure of source code or individual user data may provide little insight into how a system works or why it made a given decision, but could enable abuse or exploitation of systems, and carries significant risks to user privacy and intellectual property.

Implementation practicalities

Process is paramount in developing and offering AI technology and applications. While researchers and developers can't guarantee outcomes, they can ensure they follow basic hygiene in terms of process. Regulators can help by providing concrete guidance about expectations — for instance, the importance of carrying out risk assessments prior to launch. There are also common pitfalls to avoid, such as over-reliance on human oversight as a solution to issues presented by an AI system, or setting unworkable standards for transparency as barriers for AI use.

6. Clarify expectations for conducting risk assessments

Prior to any launch it is reasonable to expect a risk assessment to be carried out and documented, with deeper analysis of products and services that are deemed to present a higher risk. However clarification is needed as to appropriate risk thresholds to apply, and the treatment of products in the early stages of development or which receive significant updates (see **Box 2**).

Box 2

Key scoping considerations for a risk-based framework

Guidance on risk classification thresholds

Conventional approaches to assessing risk take into account the severity of harm compared against the likelihood of its occurrence. Normally severity is categorized as "catastrophic", "major", "moderate", "minor" and "negligible"; and the probability of an adverse effect as "very likely", "likely", "possible", "unlikely" and "very unlikely". Scoping the risk of an AI application in such a fashion is advised because it allows for various combinations of severity/likelihood to qualify as high-risk (e.g., not just "major/likely" but also "catastrophic/very unlikely", "minor/very likely"). Regulation should include guidance as to when the risk classification of a given AI application flips from low or medium to high, and clearly reflect that the objective is to mitigate the severity of harm, while simultaneously reducing its likelihood.

Treatment of R&D and early stage products

In the early stages of development there will often not be a clear view as to the ultimate shape of a product (indeed it may not even be clear what is technically feasible), and thus it is not possible to thoroughly assess risks or necessary consultations until a later stage. It is therefore important that confidential piloting of an AI application be allowed prior to any risk assessment, within the bounds set by existing sectoral regulation. If such pre-assessment testing is not permitted, it may result in organizations taking an unduly precautionary stance in terms of the necessary requirements and investment, which would hinder innovation.

Treatment of products which receive significant updates

Carrying out a new risk assessment should only be required when there has been a significant change to the functionality of the product that is likely to materially alter its performance in testing or safety disclosures. Generic over-the-air updates (OTAs) such as security patches, bug fixes, or simple operational improvements after placing a product on the market should not trigger a renewed risk assessment. Potential determinants of whether a modification should spark a new assessment could include: whether there has been a significant alteration in the training data or model to cater to a wider target audience; or whether a change is in response to external factors rather than to the dataset or model (e.g., if medical authorities altered the gold standard test required for a specific diagnosis).

Undertaking the upfront risk assessment should be the responsibility of those deploying the AI application, since only they know the intended context of its use. While providers of off-the-shelf, multipurpose AI component systems can provide general information about their construction and guidance on operating boundaries in foreseen use-cases, they are in no position to conduct the risk assessment because they cannot verify the end-uses to which their systems are put.

A practical approach would be for regulators to provide "due diligence" guidance, but assign responsibility for conducting the risk assessment to the organization using the AI application. The resultant documentation would provide evidence of the satisfactory completion of the risk assessment, and could be available to view on demand by regulators, or even filed confidentially with a certification body. Post launch, if concerns arose that an application had been mis-classified, remedial action could be taken via existing legal channels and sector regulators.

Box 3 outlines the key operational and organizational considerations that affect the level of risk of any given AI system, which should be part of any assessment. However, care should be taken to avoid being overly prescriptive in terms of format, to avoid inadvertently requiring exhaustive documentation of aspects posing little to no risk.

In some contexts it may be possible to adapt established design and validation processes, particularly when they stem from the same domain as the AI application in question. For example, the concept of a "failure modes, effects and criticality analysis" (FMECA), if tailored judiciously to suit the application context, may present a structured approach to documenting the expected impact of foreseeable safety risks, and the corresponding preventive measures or reactive strategies planned if such failures were to occur.

> **Example:** The European Union Aviation Safety Agency (EASA)'s AI taskforce has worked with Daedalean.ai to explore practical approaches[4] to account for neural networks in safety assessments, including carrying out a failure mode and effect analysis for ML components used in safety-critical applications.

In certain circumstances where there are legitimate concerns over possible risks relating to fundamental rights, a formal human rights impact assessment conducted by a credible expert may be warranted. This would align with Section 21 of the United Nations Guiding Principles on Business and Human Rights, which many companies, including Google, are already committed to uphold.

> **Example:** Google Cloud's AI Principles review team enlisted the nonprofit organization BSR (Business for Social Responsibility) to conduct a formal human rights assessment of their new Celebrity Recognition tool, offered within Google Cloud Vision and Video Intelligence products. BSR applied the UN's Guiding Principles on Business and Human Rights as a framework, and their assessment[5] informed not only the product's design, but also the policies around its use.

Box 3

Key considerations in assessing the risk of AI systems

What is the inherent risk of applying this technology to this specific problem?

While there may be some cases where a certain technology is inherently risky, more often the primary driver of risk is derived from the precise use context.

- It is possible to deploy an AI application in a sector that is typically seen as higher risk, in ways that are not inherently high risk — e.g., AI-enabled email systems in policing.
- Even tools that are typically seen as higher risk, like the use of facial recognition by law enforcement, can have huge variance in risk profile depending on their precise purpose — e.g., using facial recognition to authenticate officers' identities carries different risks than using facial recognition to conduct surveillance and identify criminal suspects.
- Similarly, a seemingly low risk activity such as online shopping could deploy AI in a way that is higher risk — e.g., discriminatory profiling.

How do the attributes of this particular AI system impact overall risk?

Specific design features and operational constraints and mitigations — both technological and in terms of business processes — may reduce or increase overall risk.

- **Consistency** — In some use contexts, a system that has been designed to perform reliably with a similar degree of accuracy across demographic groups or in different contexts may be deemed less risky than a system that is often more accurate but has less consistent performance across groups or operating environments.
- **Reversibility** — If operational processes are designed to allow errors made by an AI system to be easy to spot and rectify, it could reduce the risk of actual harm relative to a system that is more opaque or where mistakes cannot be easily reversed.
- **Degree of human control** — Risk assessment should reflect the difference in exposure between an AI system operating with a significant degree of human control and one that is operating with minimal human oversight. In some use contexts, autonomous systems may be deemed more risky than systems with direct human oversight. However it should not be presumed that a greater degree of human control is always lower risk since in certain circumstances the reverse may hold — for instance if the human overseeing the system had strong subconscious biases, or was impaired by fatigue or alcohol, or if the task at hand requires a degree of speed and precision that a human is unable to provide.
- **Continuous learning safeguards** — Without suitable technical constraints (e.g., a hard limit on the percentage difference in prediction permitted from that of the previous tested model), AI systems which learn and adapt in real time are likely to be more vulnerable to deliberate or unintended manipulation. With appropriate safeguards, however, continuous learning systems can provide better results in dynamic operating contexts, helping to reduce the risk posed by models whose training does not fully reflect the current operating environment.
- **Environmental mitigations** — Not all mitigations must be designed into AI systems themselves. For example, an AI application used by an organization with strong internal governance processes may pose less risk than if it were deployed by an organization lacking such stringent self-regulatory oversight. Similarly, physical barriers between autonomous robots and humans in a factory or sorting facility can ensure safety for the human workers even if the robots themselves have relatively simple safety features.

Is the overall risk of this AI system tolerable when compared to existing alternatives?

AI systems will never be perfect, but neither is human decision making. Taking into account the inherent risks of the technology and use case, and specific attributes of this particular AI system, is this risk greater or less than the risks of not using AI, and is it tolerable?

- If the risk of an AI system is less than the risk of established methods of carrying out a task, even an imperfect AI system may be preferable to continuing to rely on existing flawed methods.
- In some use contexts, it may be tolerable to accept an AI system that has higher risk and performs more poorly than a human expert carrying out the same task, if there are not enough people willing and able to do the job.

7. Take a pragmatic approach to setting disclosure standards

Google strongly endorses the notion that AI applications - especially those that pose higher risks - should be required to disclose general information about their existence and nature to those who have a legitimate interest. However it is important for regulation to be tempered by a recognition of the tradeoffs and difficulties inherent to providing detailed information about the inner workings of an AI model. High aspirations for transparency are to be encouraged, but should be clearly differentiated from the minimum acceptable threshold, which should be set to reflect the context of a particular AI application.

There are however three general principles for disclosure that should apply in all settings.

First, the organization deploying an AI application should be solely responsible for any disclosure and documentation requirements. Just as a brickmaker can't be expected to predict every possible construction their bricks could be used in, nor can a supplier of AI system components. Third parties supplying multi-purpose AI components should only be required to ensure that the terms and conditions of sale do not prevent a purchaser from meeting disclosure obligations.

Second, whenever AI is playing a substantive role in decision-making or directly interacting with people, that fact should be easily discoverable. This is particularly important in cases where people could have reasonably assumed that AI was not playing a significant part. Public disclosure will typically be appropriate for applications designed for consumer use or that make decisions affecting individuals (such as the allocation of government services or healthcare). However, information about B2B use of AI (such as in a factory setting as an aid to manufacturing or optimizing the operations or a wind farm or port) should not be required to be disclosed to those without a legitimate interest, except in rare instances where there is deemed to be a clear public interest.

Third, disclosures should be presented in clear, salient language so as to be meaningful to a wide audience, and should provide an overview of the key tasks the AI is being deployed to assist with, within the context of the application being offered. Box 4 provides additional detail on the kind of information that should be included. Where appropriate, additional technical information relating to AI system performance should also be provided for expert users and reviewers like consumer protection bodies and regulators. This could include information about how well the AI system performs for industry-standard evaluation datasets measured against key metrics; providing an indication of the frequency and cost weighting assigned to different errors (e.g. false negatives/false positives); and, if relevant, how the AI system's performance compares to existing human-performance benchmarks. with people, that fact should be easily discoverable. This is particularly important in cases where people could have reasonably assumed that AI was not playing a significant part. Public disclosure will typically be appropriate for applications designed for consumer use or that make decisions affecting individuals (such as the allocation of government services or healthcare). However, information about B2B use of AI (such as in a factory setting as an aid to manufacturing or optimizing the operations or a wind farm or port) should not be required to be disclosed to those without a legitimate interest, except in rare instances where there is deemed to be a clear public interest.

Third, disclosures should be presented in clear, salient language so as to be meaningful to a wide audience, and should provide an overview of the key tasks the AI is being deployed to assist with, within the context of the application being offered. **Box 4** provides additional detail on the kind of information that should be included. Where appropriate, additional technical information relating to AI system performance should also be provided for expert users and reviewers like consumer protection bodies and regulators. This could include information about how well the AI system performs for industry-standard evaluation datasets measured against key metrics; providing an indication of the frequency and cost weighting assigned to different errors (e.g. false negatives/false positives); and, if relevant, how the AI system's performance compares to existing human-performance benchmarks.

> **Example:** Google is investing in Model Cards[6], similar in consent to nutrition labels for food, to increase transparency and understanding around the proper use and limitations of AI models. To explore the possibilities of model cards in the real world, prototypes were published for two features of Google's Cloud Vision API, Face Detection and Object Detection. They provide simple overviews of both models' ideal forms of input, visualize some of their key limitations, and present basic performance metrics.

However, it is important for regulations to retain flexibility in the format and precise details required to be disclosed, because what is most appropriate will vary by context and

Box 4

Key disclosure requirements

Topline indication of how the AI system works

It is not appropriate to expect organizations to reveal full details about AI models or the underlying code, because this would risk undermining business confidentiality and enable adversarial gaming of the system. However, an indication should be provided as to the general logic and assumptions that underpin an AI model, particularly if they are designed for use in high risk settings. It is also good practice to highlight the inputs that are typically the most significant influences on output, as well as any inputs likely to be deemed sensitive or unexpected. Any inputs that were excluded that might otherwise have been reasonably expected to have been used (e.g., efforts made to exclude gender or race) should also be flagged.

Expectations about how an AI system will be used

When relevant, an indication should be given as to any operational expectations in mind when the system was designed, such as whether it is intended to function independently or with a level of human oversight. There is evidence that users interact with AI systems and react to errors differently depending on such assumptions, so this information will help users to build suitable mental models when they are utilising an AI application. While it is not possible to anticipate every possible use of an AI system, an indication can always be given as to the use cases in mind when it was designed (e.g., those use cases against which its performance was tested and/or for which it is being marketed).

Known limitations on performance

It will often be difficult to describe in clear lay terms the expected limitations and level of accuracy under different conditions. However, general guidance can still be given. Research has shown[10] an AI application's performance can be better contextualised by presenting it alongside existing human performance statistics where they exist. Concrete examples of successful and unsuccessful use cases are also helpful, particularly any challenging edge-cases or known pitfalls in existing non-AI approaches that the system has been explicitly designed to overcome.

audience. For example, in a narrow set of domains (e.g., medicine) where expert trust heavily depends on knowing whose decisions provided the ground truth, an indication as to the AI system's source of "ground truth" during training can help experts using the system to calibrate an appropriate level of trust, and to assess when they should rely on an AI system, and when they should instead rely on their own judgment. In other contexts this will be unnecessary.

Finally, be cautious about imposing mandatory disclosure requirements for datasets used to train an AI model, as this will seldom be practicable. Sharing data sets may conflict with copyright provisions, particularly if non-infringement is based on only temporary use of copies. It may also violate contractual obligations to not retain data supplied by business clients. Providing third party access to data could also undermine privacy, such as by requiring a central log of data to be stored or preventing data from being deleted. More fundamentally, organizations who have built products using open-source models have no reliable way to know the provenance of the data used to train the models unless the publisher has chosen to release it.

8. Workable standards for explainability and reproducibility require compromise

It seems commonsensical to expect AI systems to perform consistently and in ways that can be explained. Such attributes can help to identify harms, empower users to make informed choices, and hold AI providers accountable — all of which underpin trust and confidence in using AI systems. However, while AI systems aspire to uphold such standards (not least for the competitive advantage they imbue) governments should be judicious in how operational requirements are scoped. A sensible compromise approach is needed that reflects the practical constraints and tradeoffs imposed by different standards of explainability and reproducibility.

In the case of explainability, the problem is that explanations can be costly in terms of technical resources or trade-offs with other goals like model accuracy (e.g., if more accurate but harder-to-explain techniques must be foregone). If every outcome of an AI system were mandated to be fully traceable and supported by a detailed explanation — a far higher standard than any human-based system can meet — it would in practice restrict AI systems to an extremely limited, basic set of techniques (e.g., static decision trees). This outcome would dramatically undermine AI's social and economic benefits.

In addition, tailoring explanations to be meaningful and suit the needs of a range of audiences is difficult and time intensive[7]. The ability to trace back and explain outcomes from AI systems operating at scale on a daily basis will likely differ greatly from the more extensive probing possible during development and upfront testing. While there has been much progress in tools to support developers, such as Google's Explainable AI tool[8] for Cloud AI customers, providing explanations at scale and in real time remains challenging, in part because the detail and scope of what is needed varies significantly by sector and audience, and expectations may evolve as best practices emerge.

When it comes to reproducibility and performance consistency, the problem is one of practical interpretation. Should individual results be reproducible, or instead systemic patterns of behavior? Must all outcomes be reproducible or just certain specified ones, or certain components thereof? A too literal interpretation of reproducibility would be tricky, if not technically impossible[9], however there are some workable compromise approaches. Examples include explicit versioning of code combined with information about which data was used for testing and training models ("data lineage"), the notion of archiving snapshots, and adopting a statistical notion of reproducibility which does not require exact matching (e.g., similarly to reproducibility in scientific experiments, the expectation is not for the exact same outcome every time, but rather the same probability distribution of outcomes each time).

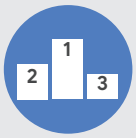
Generally there is little need for regulatory encouragement of explainability and reproducibility in AI systems, because they are already incentivised by market forces. After all, if AI systems don't provide reliably consistent outputs for the same inputs, they are unlikely to be useful and uptake will be low. Similarly, if users are unable to understand broadly how an AI system works, they are less likely to be comfortable with its use, especially in high-stakes situations. Sometimes, however, there may be practical constraints as to what's possible, and alternative approaches to delivering accountability and boosting trust must be found. One common method is to impose stricter process guardrails on an AI system's use, such as regular monitoring or even triggering human reviews in certain instances. In light of this, when crafting rules about such attributes, regulators should consciously allow room for workable compromises when necessary, identify the tradeoffs to imposing these rules, and avoid being overly prescriptive.

9. Ex-ante auditing should centre on processes

It is reasonable to expect organizations and individuals that develop, deploy and use AI systems to demonstrate that they are managing risks appropriately — particularly when the stakes are high — and to enable investigators to understand how and why systems fail in situations that lead to harm. At the same time, imposing unreasonably stringent and inflexible requirements could expose intellectual property and business confidential information, constrain innovation, and hinder the development of useful applications. A sensible balance can be struck by requiring organizations developing or using AI tools to conduct periodic audits of their governance processes using independent auditors who are professionally qualified, and entrusted to only certify organizations who meet the appropriate standards. **Box 5** highlights the key considerations for process audits, including goals, operational structures and practices, and approach to handling problems.

Box 5

Key considerations for Process Audits



Have responsible design goals and their relative prioritisation been clearly defined, and are they appropriate for the use case/application?

What constitutes responsible behavior can vary greatly across different AI applications and use contexts. In practice, individuals and organizations will face trade-offs between competing equities (e.g., security vs accessibility vs fairness) when they design, deploy and use AI systems. Organizations should be able to articulate the rationale for goals set, in a manner that allows auditors to evaluate their appropriateness.

Are robust organizational structures and practices in place to help ensure that the goals for the AI system are met?

Goals are seldom achieved without management effort and attention to ensure that technical and business decisions about AI system features are aligned. This includes having clarity over how system performance will be monitored over time, in terms of key metrics and evaluation against relevant benchmarks. Organizations should be able to describe the internal governance structures and processes in place to foster and assess progress towards the goals set over the life of the system.

Are suitable procedures in place to expose and address problems in a timely fashion, including supporting ex-post investigation of serious incidents in high risk settings?

Organizations should have clearly defined mechanisms and processes in place to monitor system performance over the life of the system, receive and review feedback from users, implementers, and third parties as appropriate, and ensure that appropriate action is taken in response to identified problems. Organizations or individuals deploying AI in high-risk settings should ensure that their systems enable thorough investigations of serious incidents, including by recording relevant input and output data for an appropriate period of time.

10. Ensure fairness benchmarks are pragmatic and reflect the wider context

When building an AI tool to assist in decision-making it is necessary to make choices upfront about how best to balance competing definitions of fairness. Different technical approaches will result in models that are equitable in different ways, and may require tradeoffs in terms of general accuracy or efficiency. Governments have a role to play in providing guidance about how to balance competing priorities and approaches to fairness, and holding organizations accountable for defining appropriate fairness benchmarks for AI systems.

In many cases, existing discrimination laws will provide a frame for balancing competing equities, particularly in highly regulated industries. For example, in most jurisdictions discrimination in lending is clearly defined in existing law, whether loan decisions are made by a human loan officer or an algorithm. Where existing discrimination laws provide clear guidelines and accountability mechanisms, new rules may be unnecessary.

But not all unfair outcomes are the result of illegal discrimination, and some AI systems may have unfair impacts in ways that are not anticipated by existing laws and regulatory frameworks. In these situations, regulators should take a nuanced approach, ensuring that organizations consider the unique historical context in which an AI system is deployed, and use appropriate performance benchmarks for different groups to ensure accountability (see **Box 6**).

Box 6

Key considerations for fairness assessments

There is no one-size-fits-all set of fairness metrics, but at a minimum regulators should ensure that organizations have clearly defined fairness goals based on historical context and clear performance benchmarks. Specifically they should be able to answer the following questions:

How have people in different groups been historically affected by this kind of product or use case? How will they be affected by this particular product?	Ideally, no groups have been historically negatively affected by this kind of product or use case, or if one or more groups have been negatively affected by this specific product or use case in the past, the particular product or use case offers clear counterbalancing benefits to these groups. This assessment should be based on significant user testing and consultations with relevant experts, including economists, sociologists, and ethicists.
How does the product perform across different user types (e.g., gender, age, skin tone, face/ body feature shapes, effect of lighting, effect of makeup/clothing, language, disabilities)?	For each application, there should be a clearly defined distribution appropriate for variant performance across groups, based on user testing, existing human levels of accuracy, published research, legal requirements, and other relevant inputs. The product should be tested across a diverse set of user groups and meet or be narrower than the target performance distribution among groups. Where appropriate, the performance distribution and/or determination process can be shared to provide users with more information and the opportunity to compare with alternative products.

11. Prioritise robustness, but tailor expectations to the context

Robustness can be thought of as combining the related notions of performance reliability, safety protections, and the ability to withstand adversarial threats. No system will be perfectly robust — what matters is providing an appropriate degree of robustness in a given context, looking across these multiple considerations and reflecting the specific nature of the risks faced and effectiveness of possible mitigation options.

As a matter of principle, all AI systems deployed in important operational settings should be built and tested for robustness. Complex systems, such as those involving interaction between multiple AI models should be reviewed as an integrated whole, rather than looking only at component models in a standalone fashion. **Box 7** highlights some key considerations when designing robust AI systems.

If testing reveals that an AI system is insufficiently robust for its intended use, mitigating action is needed. A good starting point is to consider what measures would be deemed appropriate if using a non-AI tool for the purpose and apply similar thinking, while recognising that AI may introduce new kinds of risk, or significantly heighten or lower certain risks, relative to non-AI approaches.

Any precautions imposed should be in proportion to the harm that could ensue and the viability and likely effectiveness of the preventative steps proposed. It is important that flexibility is retained so that actions can be tailored to suit the context. Mandating particular techniques legislatively may inadvertently undermine longer term robustness by discouraging organizations from developing improved techniques and approaches.

Key considerations in designing robust AI systems

Box 7

Design for failure

The ability to withstand attack and "fail gracefully" is crucial, but more generally robustness can be interpreted as affirmatively and intentionally designing an AI system to cope with failure and adapt to new situations. For example:

- Coding in hard constraints to prohibit unexpected system behaviours outside of the range deemed safe. Adding such constraints needs to be done judiciously so as to not undermine the system's resiliency in adapting to new situations.
- Formal pre- and post-launch vulnerability testing processes, as well as processes to support monitoring throughout the life of an AI system. No system will ever be perfect, and most failures that occur will be unexpected.

Tailor safeguards to suit the context

In settings where mistakes or attacks could have extreme consequences and be hard to reverse, it may be necessary to apply stringent guardrails that prevent the system from operating if inputs or outputs fall outside a predefined "safe" range. In other situations, stopping an automated process entirely or handing off control to a human operator on the fly could create more serious risks. In those cases it may be more appropriate to prioritise checking for anomalies and errors early and having established processes to remediate.

Be willing to pull back

Companies and developers must think carefully about the likelihood and consequences of problems their AI system may face, including threats posed 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 an AI application until better protection mechanisms are available.

12. Be wary of over-reliance on human oversight as a solution to AI issues

Human input is central to an AI system's development. From problem and goal articulation, to data collection and curation, and model and product design, people are the engine driving AI innovation. Even with advanced AI systems able to design learning architectures or generate new ideas, the choice of which to pursue must be guided by human collaborators, not least to ensure that these choices conform to legal and financial constraints. Similarly, people play a vital role in the verification and monitoring of a system, such as choosing which tests to run, reviewing results, and deciding if the model satisfies the performance criteria required to enter or remain in real-world use. And of course, human users provide essential feedback to improve AI systems over time.

However, regulators should be cautious in mandating human oversight. Forms of oversight that are commonsensical in one setting will be harmful and undermine the core essence of an AI application in another. For example, requiring an AI system's output to be reviewed by a person before being acted upon may make sense for some applications (e.g., AI systems used for critical, non-time-sensitive medical diagnostics). However, for other applications, it could lead to sluggish output, reduced privacy (if it means more people see sensitive data), or undermine accuracy (if human reviewers lacked the necessary expertise or were more biased). At an extreme, it could even put people at risk, for example by delaying automated safety overrides.

In addition, wider practicalities of implementation need to be considered. For instance, in contexts where a human review of an AI system's recommendation is offered, there must be reasonable bounds put on the timeframe for appeals. Similarly, it's important to ensure that people who are tasked with reviewing an AI system's output are thoroughly trained and have a deep understanding of the AI's capabilities and limitations.

Ultimately, AI systems and humans have different strengths and weaknesses. In many contexts, it is possible that a team of humans and machines will perform better than either acting alone. 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 involving a human would increase the risk of mistakes. Similarly, while a lot of attention has focused on the risk that poorly designed and applied AI systems might have baked-in unfair bias, even the most well-intentioned people are vulnerable to implicit bias in their decisions. This is not to imply that there is no problem with unfairly 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. Selecting the most prudent combination and form of human oversight comes down to a holistic assessment of how best to ensure that an acceptable decision is made, given the circumstances.

In closing

While Google sees AI as a vital tool to help solve the world's most pressing problems—from improving information quality, to boosting health and accessibility, to aiding scientific breakthroughs—we are not blind to its challenges. The AI ecosystem will only thrive if it is developed and deployed responsibly. We are committed to playing our part, but cannot do it alone and there is an important role for governments to play. As our CEO Sundar Pichai put it, AI is too important not to regulate, the only question is how.

This document highlights practical considerations in providing robust protections and oversight tailored to suit the myriad of contexts in which AI is deployed, while allowing the innovation needed to deliver on AI's promise. We hope that it is helpful to regulators as they seek to translate new and long-standing legislative and regulatory principles into practical requirements and expectations for responsible AI development.

End notes

1 Google's AI principles were published in June 2018, available at https://ai.google/principles/

2 "Perspectives on issues in AI governance" was published in Jan 2019, available at https://bit.ly/2Rb3sX1. It called on governments to provide concrete guidance in 5 key areas (explainability standards, fairness appraisal, safety considerations, human-AI collaboration, and liability frameworks)

3 Further information on Google's perspectives on AI policy considerations, including links to our public submissions to government consultations, is available at http://ai.google/responsibilities/policy-perspectives

4 https://www.easa.europa.eu/document-library/general-publications/concepts-design-assurance-neural-networks-codann

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

6 https://modelcards.withgoogle.com/about

7 An added complication is that the same audience may expect different kinds of explanations (or may not even want one at all) in different contexts. For more details on the complexities and challenges of explainability, see pp 8-12 in https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf

8 https://cloud.google.com/blog/products/ai-machine-learning/google-cloud-ai-explanations-to-increase-fairness-responsibility-and-trust

9 There are very good reasons why AI systems may be designed in a manner that they do not provide the identical output for the same input. For instance, systems which use differential privacy have built-in randomisation, with random noise carefully injected in order to protect individual privacy. More generally, using the same training data will not necessarily yield models with the same learnings due to the nature of the techniques involved. For example, stochastic gradient descent (SGD) is one of the most effective and state-of-the-art techniques for machine learning and involves estimating objective function gradients on random subsets of the training dataset. The reason for this is to reduce the computational burden, thus allowing for faster iteration, which speeds up the learning process. However, because the model's training is based on random subsets of the training data, it is not possible to guarantee that the same training data would lead to the same model output. Privacy safeguards can also inhibit storing of necessary data. For example, AI systems that provide video or audio recommendations are updated over time, changing in response to the availability of content and user reactions. The only way such systems could be precisely replicable over time would be if every interaction of every user was stored indefinitely, which would be unacceptable from a privacy point of view.

10 "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making (Nov 2019); https://dl.acm.org/doi/10.1145/3359206

11 "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making (Nov 2019); https://dl.acm.org/doi/10.1145/3359206

Google