Consultation on the white paper on AI - a European approach

Google's submission
May 28, 2020

Table of contents

Overview

Google welcomes the publication of the European Commission's White Paper on AI. As the White Paper notes, AI offers many benefits for citizens and the economy, and its potential to improve lives is profound — as evidenced by the many ways that AI is being used in the context of current COVID-19 pandemic.

At Google, AI is used to make products more useful - from email that's filtered for spam and easier to compose, to a digital assistant you can speak to naturally; as well as to help tackle urgent problems as we currently see during the coronavirus crisis. To give just two examples: Google researchers have shown AI can help doctors spot breast cancer in mammograms more accurately; as well as help to fight climate change by making hyperlocal forecasts of rainfall more quickly and precisely than existing tools. At the same time, Google is working to address pressing concerns about the potential negative consequences of AI, from spotting deepfakes to combating nefarious uses of facial recognition. There will inevitably be more challenges ahead, and Google is committed to engaging constructively to help address them, such as through our membership in the EU High-Level Expert Group on AI, the OECD's AI policy observatory, and other global fora.

This document provides further commentary on the White Paper, expanding on Google's responses to the consultation survey. Topline points from each section are as follows:

- **Ecosystem of excellence**
  Europe is well-positioned to play a leading role in AI research and application given its world-class universities and thriving developer community. In Google's view, it is particularly important for the Commission to focus on improving digital skills, developing AI expertise among SMEs and encouraging governments to promote public sector adoption of AI to help create a thriving ecosystem. In parallel, there is a need to reduce fragmentation in the R&D landscape. Google supports the Commission's proposal to create a lighthouse centre for AI research, innovation and expertise, although it is vital to strike the right balance in the degree of centralisation.

- **Ecosystem of trust**
  Smart government approaches to regulation will play an important role in building trust and ensuring that AI is used responsibly while encouraging innovation. Google supports the Commission's goal to achieve trustworthy, human-centric AI, and agree conceptually with the need for a well-defined risk-based approach to AI regulation that doesn't apply a "one size fits all" framework across AI's myriad applications. However, Google believes that the definition of high risk AI applications proposed in the White Paper requires further clarification to ensure regulation is targeted at the right use cases, provides legal certainty and does not discourage the development of AI. Most crucially, it is important that a more proportionate, risk-based approach is taken: balancing potential harms with the many social and economic benefits promised by AI, and by more clearly acknowledging the opportunity costs of not using AI in a specific situation.

The White Paper outlines a set of suggested mandatory requirements, several of which Google believes could significantly hamper the development and availability of socially and economically beneficial AI applications. Among the most concerning are requirements relating to training data, which could present significant practical challenges, conflict with existing laws (such as GDPR and copyright), and hinder using AI quickly and effectively to respond to crises such as the current COVID-19 pandemic. Another substantial concern is the notion of requiring reproducibility, which if defined literally would be technically impossible for many AI systems. In general, Google urges the Commission to work closely with practitioners when crafting and clarifying such requirements to ensure that what is proposed is technically feasible and appropriate to meeting the Commission's goals.

With respect to compliance and enforcement, we recommend expanding established due diligence and regulatory review processes to include the assessment of AI applications. This would avoid unnecessary duplication of efforts and likely speed up implementation. For the (probably) rare instances when high-risk applications of AI are not obviously covered by existing regulations, we would encourage clear guidance on the "due diligence" criteria companies should use in their development processes. This would enable robust upfront self-assessment and documentation of any risks and their mitigations, and could also include further scrutiny after launch.

- **Safety and liability**
  Safety and liability frameworks must provide users of AI applications with sufficient protection, so if significant shortcomings are identified they must be addressed. However the Commission's report appears to conflate the notion of health and safety with concepts which fall outside the scope of product safety (e.g., cyber security, ethics, privacy and mental health). Any review of EU safety regulation should focus exclusively on areas where the unique properties of AI, IoT or robotics create a risk to the health and safety of consumers. With regard to liability, the Commission should be wary of expanding the scope of the existing liability framework to cover AI software and services, as such a dramatic and unprecedented expansion would constrain innovation and disproportionately hinder European SMEs. In Google's view, the current Liability framework remains fit for purpose, being both effective and technology neutral, so sweeping changes are not needed. Any contemplated changes should be supported by clear evidence, and a strong consensus among legal experts, that the current framework is inadequate.

More detail on all of these points can be found below, and Google would be delighted to answer any further questions the Commission might have.

Section 1 - An ecosystem of excellence

Google strongly believes in the positive contribution AI makes in Europe, including the many economic, social, and safety benefits it will create for European businesses of all sizes, civil society, and individual citizens. The current COVID-19 situation offers a vivid demonstration of the contribution that AI tools can make. In the short term AI is being used to boost knowledge-sharing, enable better prediction of health trends, and contribute to research for a cure[1]. In the longer term, AI can be a vital aid for business recovery, helping companies to more rapidly scale up and boost their productivity.

Google has been prioritising investment in advanced technologies such as AI and machine learning. These technologies make Google's core products and services much more useful to the public, including Android, Assistant, Cloud, Gmail, Maps, Photos, Pixel, Search, YouTube, and many more. Google is also creating tools to ensure that everyone can access AI, including researchers and developers, entrepreneurs and businesses of all sizes, academics, nonprofits, and governments. Wider accessibility and trust is how AI will have its biggest impact and how society can reap its full promise. With a thriving developer community, and world-class universities, Europe is well positioned to play a leading role in AI research and application and Google looks forward to playing our part in this.

The Commission has rightly identified a need for focus on investment of AI by European citizens, businesses and the public sector to ensure that the sector continues to grow. Section 1 of the Consultation is focused on how an ecosystem of excellence which supports the development and uptake of AI across the EU economy can be built. It sets out six actions to help ensure an ecosystem of excellence, all of which are important. In terms of relative priorities, Google believes the Commission should centre its efforts on: supporting the research and innovation community; ensuring the right skills are in place so all are able to prosper from the benefits AI can bring; supporting SMEs by raising awareness about the potential benefits of AI and promoting the knowledge transfer and development of AI expertise among SMEs; and promoting the adoption of AI by the public sector. Further detail on Google's work in this area follows, as well as thoughts on how the Commission can better support these areas.

[1] The Council of Europe have provided a helpful roundup of some of the many ways that AI is being used to help tackle the Covid-19 pandemic:
https://www.coe.int/en/web/artificial-intelligence/ai-and-control-of-covid-19-coronavirus

In your opinion, how important are the six actions proposed in section 4 of the White Paper? (Choose from 1-5: 1 is not important at all, 2 not important, 3 neutral, 4 important, 5 is very important OR No opinion)

Working with Member states	4
Focussing the efforts of the research and innovation community	5
Skills	5
Focus on SMEs	5
Partnership with the private sector	4
Promoting the adoption of AI by the public sector	5

Are there other actions that should be considered? (500 characters max)

The Commission has identified a good combination of actions to ensure the development of an ecosystem of excellence. Google supports the Commission's intention to focus on uptake and deployment of AI by European citizens, businesses, researchers and the public sector. This is even more vital in the current context, since AI and digital technologies will be critical components of the economic recovery.

In your opinion, how important is it in each of these areas to align policies and strengthen coordination as described in section 4.A of the White Paper?
(Choose from 1-5: 1 is not important at all, 2 not important, 3 neutral, 4 important, 5 is very important OR No opinion)

Strengthen excellence in research	4
Establish world-reference testing facilities for AI	3
Promote the uptake of AI by business and the public sector	5
Increase the financing for start-ups innovating in AI	4
Develop skills for AI and adapt existing training programmes	5
Build up the European data space	3

Are there other areas that should be considered? (500 characters max)

Supporting organisations of all sizes to develop the skills necessary to make responsible and effective use of AI is vital. In parallel, there is a similar educational challenge to equip those in the wider ecosystem who will play crucial supporting roles in guiding the uptake of AI. To support this latter challenge, Google is offering workshops for policymakers new to the field to learn the basics of AI and machine learning from expert practitioners.

Research and development

With a thriving developer community and world-class universities, Europe is well positioned to play a leading role in AI research, development and application. However, as the White Paper identifies, the fragmented landscape of existing centres of competence has resulted in research efforts that lack the coordination, resources and scale to compete globally. The Commission's proposal to address this by creating a lighthouse centre for AI research, innovation and expertise in Europe is an excellent initiative that Google would be happy to support.

In Google's view, harbouring world-leading research and innovation is key to realising the potential of AI and to ensuring that cutting edge AI continues to be developed across Europe. Google helps to support this ambition in a number of ways:

- **Google has multiple AI research centres in Europe**. For example, Google's first European research hub in Zurich is now a leading contributor to Google's progress in machine perception and language understanding. In Berlin and Amsterdam, the teams work on a range of topics from foundational to more applied research involving data comprising text, images, video, audio and more. In Paris, the team covers a variety of areas within computer science and mathematics, including algorithmic foundations and theoretical underpinnings of deep learning to operations research, mathematical programming, reinforcement learning, natural language processing, machine perception, data compression and computational biology. There are also multidisciplinary researchers from Google's People and AI research unit (PAIR) now embedded with engineering teams in London and Paris. Having a permanent physical base in European locations makes it possible for research teams to engage with the local AI community, such as hosting regular events and making it possible for Google's engineers to teach part-time at local universities.

- **Making Google's AI advances accessible to everyone:** Google is committed to collaborating with the wider AI community, including researchers and developers, entrepreneurs and businesses of all sizes, academics, nonprofits, and governments. Critical to this approach is open-sourcing of AI tools such as TensorFlow, a platform that makes machine learning faster, smarter and more flexible. Google researchers also openly publish and share their work at conferences, often accompanied by the supporting datasets and models. So far, Google researchers have open-sourced more than 60 datasets useful for training AI models, and created a Dataset Search tool to help developers track down other publicly available datasets. Google Cloud brings this technology to the enterprise world, offering a range of AI-powered products and solutions, from pre-built APIs (i.e., building blocks for using AI to tackle tasks related to sight, language, conversation, and structured data), through to end-to-end solutions that are helping to transform sectors such as financial services, retail, healthcare, and beyond. Google partners with many European companies to boost their AI capabilities, including Siemens, Lufthansa, Airbus, and more.

- **Partnerships with European AI researchers:** Google is committed to Europe and to supporting the continued growth of the European AI ecosystem. The scale of Google's academic and technical partnerships ranges from institutional level research agreements (e.g., with INRIA - National Institute for Research in Digital Science and Technology in France, TUM - Technical University of Munich in Germany), through to project-based contracts or grants to individual researchers (e.g., via Google's visiting researcher program or PhD fellowships). The ambition is to have an open arena for visiting researchers and collaborations across fields of mutual interest. For instance Google's partnership with TUM is oriented on automation and industrial robotics; whereas the partnership with INRIA began with a focus on computer vision. Across Europe, Google has hosted over 30 visiting researchers and supported dozens of PhD students in support of AI-related academic research. Google also participates in more targeted knowledge-sharing initiatives, such as our recent submission to the Commission's repository of AI projects helping to tackle the COVID-19 situation. This includes the Kaggle COVID-19 Open Research Dataset Challenge, which hosts an open dataset of over 134,000 scholarly articles with a call to action for the world's AI experts to develop text and data mining tools that can help the medical community develop answers to high priority scientific questions.

In your opinion how important are the three actions proposed in sections 4.B, 4.C and 4.E of the White Paper on AI?
(Choose from 1-5: 1 is not important at all, 2 not important, 3 neutral, 4 important, 5 is very important OR No opinion)

Support the establishment of a lighthouse research centre that is world class and able to attract the best minds	5
Network of existing AI research excellence centres	4
Set up a public-private partnership for industrial research	3

Are there any other actions to strengthen the research and innovation community that should be given a priority? (500 characters max)

There is currently a fragmented landscape of AI research centres in Europe. Streamlining and strengthening coordination between them would boost synergies and partnership opportunities. The proposal to create a lighthouse centre for AI research in Europe is an excellent initiative that could help, although it is vital to strike the right balance in the degree of centralisation. A single institute spread across several locations could be a good model.

Skills

The White Paper rightly identifies the need to underpin its approach to AI by a strong focus on skills in order to fill competence shortages. Google welcomes the aims of the updated Digital Education Action Plan to help make better use of AI-based technologies such as learning and predictive analytics with the aim to improve education and training systems and make them fit for the digital age. Ensuring people are able to improve their digital skills is vital to Google and we have a number of initiatives to support this. Most notably:

- **Grow with Google:** Established in 2015, the Grow with Google (GwG) programme is designed to equip people and businesses with the digital skills needed for the future workplace. The curriculum includes a dedicated module on understanding the basics of AI. GwG is having a demonstrable impact on jobs and business growth across Europe, with 6.7 million Europeans participating to date, 29% of whom report having found a job or grown their business or career as a result of the training. In 2020, the ambition is for another 1 million Europeans to be trained in digital skills via the GwG programme.

- **Learn with Google AI:** Google's Learn with Google AI programme provides deeper training for people who want to develop machine learning skills, targeted at beginners through to seasoned practitioners. Materials available include a Machine Learning Crash Course, which was designed by Google to train over 20,000 of our own engineers (now available in 11 languages including English, French, German, Spanish and Russian).

Focus on SMEs

Promoting knowledge transfer and supporting the development of AI expertise for SMEs is vital. SMEs are the backbone of Europe's economy, representing 99% of EU businesses, two thirds of private sector employment and accounting for over 85% of new jobs in the past 5 years. Driving knowledge and uptake of AI among SMEs will allow these businesses and their customers to reap the many benefits promised by AI applications.

In addition to the skill-building programmes referenced earlier (Grow with Google, Learn with Google AI), there are a number of products and initiatives that are designed to encourage and assist SMEs in the savvy adoption of AI. These include:

- **Cloud AutoML:** This is a suite of products that enable the training of high-quality machine learning models with minimal effort and machine learning expertise. Using Cloud AutoML, SMEs do not need to have AI expert engineers in-house, as the program allows their existing developer team to custom tailor models to their specific business needs. Cloud AutoML is now available to create models that perform a wide variety of tasks, including natural language processing, translation, analysis of images (including still imagery and video), as well as working with general tabular data.

- **Machine Learning Checkup:** AI is a powerful technology, but it is not always the most appropriate solution. It is vital that businesses understand when AI is a sensible choice, so that they can target their resources and avoid common pitfalls. That's why in 2019 Google partnered with the School of Management at Milan Polytechnic to create a free tool to help companies understand the potential benefit of AI for their business. After filling in an online questionnaire, companies receive a customised report that highlights specific applications of AI that could be useful, as well as the preparation necessary to implement. In January Google announced the Machine Learning Checkup tool will be extended to 11 markets across Europe.

In your opinion, how important are each of these tasks of the specialised Digital Innovation Hubs mentioned in section 4.D of the White Paper in relation to SMEs?
(Choose from 1-5: 1 is not important at all, 2 not important, 3 neutral, 4 important, 5 is very important OR No opinion)

Help to raise SME's awareness about potential benefits of AI	5
Provide access to testing and reference facilities	3
Promote knowledge transfer and support the development of AI expertise for SMEs	5
Support partnerships between SMEs, larger enterprises and academia around AI projects	4
Provide information about equity financing for AI startups	3

Are there any other tasks that you consider important for specialised Digital Innovation Hubs? (500 characters max)

N/A

AI adoption in the public sector

As the White Paper points out, it is important to increase uptake of AI in the public sector given the potential benefits that it offers.

However, in addition to the sector focus areas identified by the Commission, Google believes there are opportunities for governments themselves to benefit from the use of AI in their operations. Most people expect to receive the same level of service from the government as from private companies. Harnessed well, AI has the potential to help public sector agencies respond faster and with greater nuance to citizen queries. In addition, by showcasing how AI can be practically and sensitively applied, government could help to lead the way as role models.

Of course, doing this in practice will be challenging. Government faces the same uncertainties as business when it comes to determining the most appropriate manner in which to deploy AI. Google recommends that public procurement of AI should be tethered to specific problems (rather than procuring technology for technology's sake) for it to have maximum impact. It may also be helpful for the Commission to provide basic principles-based guidance for government agencies when considering funding AI solutions, such as are currently in development by the UK's Office for AI in partnership with the World Economic Forum.

More specifically In the context of the current COVID-19 situation, Google has submitted Google Cloud Rapid Response Virtual Agent, which helps governments, healthcare organisations, and other businesses quickly build and deploy a customised Contact Center AI virtual agent that can help serve their customers in 23 languages who are looking for accurate and current information during the COVID-19 pandemic.

Section 2 - An ecosystem of trust

AI offers many economic, social and safety benefits for European businesses of all sizes, civil society and individual citizens. However, such a powerful technology raises equally powerful questions about its use, including the best way to build fairness, explainability, privacy and security into AI systems, and how to make the unavoidable trade-offs that are necessary to build trustworthy AI systems. Without a foundation of trust, the opportunities that AI offers will not be fully realised.

Google welcomes the current discussions on how to create the proper framework for trustworthy AI in Europe and is committed to engaging actively and constructively in the multi-stakeholder European discussions about AI governance, including through the EU High-Level Expert Group on AI. Boosting public trust in AI is a core objective that unites technologists, businesses, policymakers and citizens.

It is in this spirit that Google offers the following feedback on the concerns raised and governance proposals included in Section 2 of the White Paper. This provides more context for the answers provided to the survey questions, as well as additional feedback on the suggested definitions of "AI" and "high risk" which are so crucial to providing clarity on the scope of the proposed regulation.

Do you think that the concerns expressed above can be addressed by applicable EU legislation? If not, do you think that there should be specific new rules for AI systems?

Current legislation is fully sufficient
Current legislation may have some gaps
There is a need for a new legislation
Other
No opinion

There are already many regulations and legal codes that are technology neutral in nature, and thus broad enough to apply to AI, but it is worth evaluating if there are gaps in the context of specific concrete problems. Any gaps identified should be addressed via practical, principles-based rules which build on existing legislation, so as to avoid creating overly complex or conflicting legal obligations.

Do you have any other concerns about AI that are not mentioned above?
Please specify (500 characters max)

Google is concerned that the opportunity cost of not using AI is not sufficiently reflected in policy debates. When considering the risks of AI, it is vital to acknowledge there are also flaws in existing (non AI) approaches. We should compare the risks of using AI systems against existing approaches. If an imperfect AI system were shown to perform more accurately than the status quo at a crucial life-saving task, it may be irresponsible to not use the AI system.

In addition to the existing EU legislation, in particular the data protection framework, including the General Data Protection Regulation and the Law Enforcement Directive, or, where relevant, the new possibly mandatory requirements foreseen above (see question above), do you think that the use of remote biometric identification systems (e.g. face recognition) and other technologies which may be used in public spaces need to be subject to further EU-level guidelines or regulation?

No further guidelines or regulations are needed
Biometric identification systems should be allowed in publicly accessible spaces only in certain cases or if certain conditions are fulfilled (please specify)
Other special requirements in addition to those mentioned in the question above should be imposed (please specify)
Use of Biometric identification systems in publicly accessible spaces, by way of exception to the current general prohibition, should not take place until a specific guideline or legislation at EU level is in place
Biometric identification systems should never be allowed in publicly accessible spaces
No opinion

Please specify (if relevant)

Ultimately it is up to governments to decide on particular approaches to further regulation of these technologies. However, some important factors that governments should consider include: whether these technologies are required for public security; if they have been pre-approved as being reasonable and proportionate use; and whether there is a practical way of achieving the same ends without the use of such sensitive data.

Do you believe that a voluntary labelling system (Section 5.G of the White Paper) would be useful for AI systems that are not considered high-risk in addition to existing legislation?

Very much
Much
Rather not
Not at all
No opinion

Do you have any further suggestion on a voluntary labelling system? (500 characters max)

A labelling system risks placing a significant burden on SMEs to comply. This would favour large players who can afford to meet the requirements whilst delivering minimal benefit to consumers. There needs to be broad agreement on standards before such a scheme could be feasible or helpful. Given the pace of change, any scheme would have to be very flexible to work as intended. Existing self-regulatory approaches such as Google's AI principles should also be taken into account.

Scope of future EU regulatory framework

Definition of AI

A clear and widely understood definition of AI will be a critical foundational element for an effective AI regulatory framework. Google acknowledges the challenge in crafting an acceptable definition that remains germane over time, particularly in light of the diverse opinions and lack of consensus among industry and academic experts.

The White Paper describes the main elements that compose AI as "data" and "algorithms." Such a broad framing effectively puts all contemporary software potentially in scope. A narrower definition is needed to avoid over-regulation and to focus on the subcategory of AI systems where important policy issues are most likely to arise.

In this regard, it will be important to reflect the clear line between the latest wave of AI systems that learn from data and experience, and traditional software and AI-based control systems that operate according to hard coded rules, which have long been embedded in a wide variety of high-risk systems (from flight control to pacemakers to industrial settings). Traditional systems are predictable because they perform the same way when presented with the same circumstances, according to the rules they were given — even if past experience has shown that to be sub-optimal. In contrast, modern AI systems learn from experience, either from training data supplied or from data gathered in use, and thus the way they behave in any given circumstance is more difficult to anticipate. The risks associated with traditional software and control systems are already adequately addressed by existing regulation; it is only modern AI systems with the ability to learn from experience that may present additional risks.

Two possible definitions for AI are referenced in the White Paper: the first published by the Commission in its Communication on AI for Europe[2]; and the second the subsequent, derivative refinement from the High Level Expert Group[3]. While Google believes that neither is ideal for legislative purposes, we would propose the following underlined alterations to the HLEG definition. This narrows the focus to systems able to learn from experience, excluding traditional rule-based AI, and removes the reference to hardware, which cannot behave intelligently without software underpinning it:

> *Artificial intelligence (AI) systems are software ~~(and possibly also hardware)~~ systems designed by humans that, given a complex goal, are taught by their designers or learn from experience how to act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or*

[2] COM(2018) 237 final, p. 1: "Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)."

[3] High Level Expert Group, A definition of AI, p. 8: "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions."

unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. ~~AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.~~

Definition of high risk

The White Paper proposes that an AI application should be considered "high-risk" if both the sector and the specific intended use involve significant risk. While Google is supportive of the risk-based approach conceptually, as outlined by Google's CEO Sundar Pichai[4], a number of adjustments are needed to ensure that any potential regulation is targeted at the right use cases, provides adequate legal certainty, and does not discourage the responsible development and diffusion of AI.

If you think that new rules are necessary for AI systems, do you agree that the introduction of new compulsory requirements should be limited to high-risk applications (where the possible harm caused by the AI system is particularly high)?
Yes | No | Other | No opinion
If other: (500 characters max)

Conceptually, Google supports a risk-based approach to a new regulatory framework but it is important to ensure that any potential regulation is targeted at the right use cases, provides legal certainty and does not discourage the responsible development and diffusion of AI. The Commission must be clear in its risk assessments that it is taking into account the likelihood of harm and not just the severity of the harm, as well as a nuanced consideration of the opportunity cost of not using AI.

Do you agree with the approach to determine "high-risk" AI applications proposed in Section 5.B of the White Paper?
Yes | No | Other | No opinion
If other: (500 characters max)

More nuance and proportionality should be added to the risk assessment criteria to make it easier for companies to understand when their technology may fall into this category. To do this, the Commission should better reflect well-established interpretations of risk, reflect wider operational context when assessing risk, and factor in the opportunity cost of not using AI in the definition. The "exceptional instances" clause is too open-ended and should be removed as it creates legal uncertainty.

If you wish, please indicate the AI application or use that is most concerning ("high-risk") from your perspective: (500 characters max)

N/A

[4] Sundar Pichai, Why Google thinks we need to regulate AI, FT 19 January 2020: https://www.ft.com/content/3467659a-386d-11ea-ac3c-f68c10993b04

1. **Remove the "exceptional instances" clause**

The notion of two cumulative criteria — an exhaustive list of sectors, as well as clarity over what constitutes a high risk use of AI within them — is, broadly, a workable approach to defining high risk applications. However the "exceptional instances" clause is too open-ended. Google would recommend this clause be removed in its entirety because, as written, it does not provide sufficient clarity for companies to have confidence that any specific application is in or out of scope. With some creativity in defining high risk sectors to also include identified high risk functional areas, the illustrations of 'exceptional instances' could be easily encapsulated within the enumerated high risk sectors.

For instance, adding Recruitment and Employment as a functional sector could cover specific high risk uses of AI in recruitment and hiring or in situations impacting worker's rights. Similarly, adding Security and Monitoring as a functional sector would cover certain high risk uses of AI for the purpose of remote biometric ID or other intrusive surveillance. While the vast majority of AI use in such functional sectors will be innocuous and low risk (equivalent to the use of AI in hospital scheduling referenced in the White Paper), the low risk instances should be ruled out of scope by the second prong of the high risk test.

In addition, unqualified references in the White Paper to AI applications affecting consumer rights as potentially being in the high-risk category seem overbroad, unjustified and counterproductive to the objective of focusing only on well defined areas of high risk. Also, should a specific AI application in a high-risk sector pose substantial risks to consumer rights this would still be caught by the general approach (specifically, prong two of the test).

2. **Add more nuance and proportionality to the risk assessment criteria**

While Google agrees with the spirit of the proposed criteria, several changes are needed to ensure adequate clarity, focus and predictability.

- **More closely align the definition with existing operational interpretations of risk:** The White Paper takes a binary approach with only "high-risk" AI applications falling within the scope of the proposed regulation, which is overly simplistic. Conventional approaches to assessing risk are more nuanced, taking into account the severity of harm compared with the likelihood of its occurrence. Normally severity is categorised 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 would mean that various combinations of severity/likelihood could qualify as high-risk (e.g., not just "major/likely" but also "catastrophic/very unlikely", "minor/very likely"). The White Paper touches on this by referencing the targeting of sectors where significant human-consequential risks "can be expected" and where an AI application is used in such a manner that "significant risks are likely". However, to ensure proportionality, the definition should be augmented to better reflect well-established interpretations of risk as a function of severity and likelihood, and to provide more guidance as to when the risk classification of a given application would flip from low or medium to high. More clearly reflecting a nuanced understanding of high-risk within the framework

would also make clear that the objective of the framework is to mitigate the severity of harm, while simultaneously reducing its likelihood.

- **Factor in the opportunity cost of not using AI:** the definition should reflect the potential for risk substitution, particularly in situations where the status quo (in which AI is not used) poses significant danger. In instances where the alternative of not using AI poses greater risk than the risk posed by deploying an AI system, it will be important for the regulatory framework to consider this carefully in the risk assessment so as not discourage AI's net beneficial use.

- **Clarify reference to immaterial damages and align with PLD wording:** The Product Liability Directive (PLD) defines damage as death, personal injury or damage to property. For consistency Google suggests to use similar wording in the high risk definition. As it stands, the term "immaterial damage" is not a known legal concept and should be clarified. As written it could mean anything from economic loss to hurt emotions, and could lead to legal uncertainty, discouraging investment and innovation. In spirit it is also largely covered by other laws (e.g., those relating to data protection and privacy, non-discrimination, defamation and freedom of expression) and its inclusion would be excessive and could have unintended consequences. For example, YouTube's monetisation classifiers use AI to automatically demonetise a YouTube creator's content where it is in breach of guidelines, for instance when it spreads hate speech. As written, such an action could be interpreted as producing significant effects for the rights of an individual or immaterial damage (in the form of economic loss or mere moral or reputational effects on a person). However, Google believes considering such actions high risk would not be in line with the spirit of the proposed framework. To be clear, this is not to suggest that there are not certain immaterial damages which may be relevant to consider within the high risk framing (e.g., the right to non-discrimination) — however these should be explicitly listed to avoid over-broad scoping.

- **Reflect wider operational context when assessing the level of risk:** Organisations using AI will have more encouragement to invest in additional mitigations and safeguards to reduce risks if doing so reduces the regulatory burden. There are numerous operational and organisational considerations that affect the level of risk in any given instance. The definition should support more nuanced assessment to reflect such factors. In particular:

    - *Internal governance structure:* The extent and effectiveness of an organisation's internal governance processes will have a significant impact on the level of risk that is posed. Any AI application used in an environment with strong internal oversight (such as at Google) will pose less risk than if it were being developed by an organisation without such stringent self-regulatory processes already in place.

    - *The degree of human control:* It is not clear whether the framework would consider an AI system operating with a significant degree of human control in a high-risk environment less risky than one that is operating with minimal oversight. Nor whether it would reflect the nuance that in some circumstances

the reverse would be true (e.g., if the human overseeing the system had strong subconscious biases, or was otherwise impaired such as by fatigue, drugs/alcohol, distraction, or relevant physical disability).

- *The extent to which an AI-based decision can be reversed:* If an error created by an AI system is easy to spot and review, it could be considered lower-risk due to the extent to which the error is easily rectified, even if it would have otherwise caused harm.

- *The nature and purpose of the AI application:* It is possible to deploy an AI application in a high-risk general activity such as policing, but incorporate it in ways that are not inherently high risk - such as AI-enabled email systems. Similarly, a seemingly low-risk activity such as online shopping could deploy AI in a way that is high-risk - such as discriminatory profiling. In assessing the risk presented by introducing AI it is thus not adequate to look solely at the domain of use - risk assessment must stem from understanding the precise context of the specific use of AI.

- *Differentiation in risk due to technical design elements:* In practice there are many different AI techniques, each with different strengths and challenges. The choice of which technique to use and the processes put in place to support it could have a significant effect on the level of risk.
    - For example, consider online[5] vs offline[6] learning systems. In offline training, a system can be rigorously tested before deployment and you can be confident that it won't change its range of predictions when it goes live. In contrast, online learning systems (which in practice are rare) by their very nature adapt to the training environment in real time, and are thus more vulnerable to deliberate or unintended manipulation by the training environment. However, this is not to suggest that offline trained systems should always be considered lower risk. Just because testing is possible does not mean that it is always carried out thoroughly, especially in time-sensitive contexts where rapid model updates are required, and even well-tested offline learning systems can produce undesirable outcomes when they encounter new situations in the wild.

[5] Online learning refers to an AI system which takes in new inputs sequentially as they arise and ingests those values into a processing system. The learning system evaluates these new values and updates its model of the environment in real-time to improve the accuracy of the predictions it generates. These types of systems are typically used to train algorithms in environments where there is too much training data to process at once or where the data appears as a function of time. The algorithm can instead dynamically adapt to the changing environment and provide novel predictions which emerge through live interaction with the problem space.

[6] Offline learning - sometimes known as batch learning - takes into account all possible training data available to the system before testing. During this pre-test phase the algorithm is trained to achieve some objective given available training data and is then deployed in a test or live environment on novel test examples that have not been seen before. Once it is live the algorithm is unable to change its model of the world and so cannot update its behaviour or predictions in light of new test examples or changing state of the environment. This is not to suggest the model is never updated however. In practice, engineers rapidly cycle through different models, repeatedly fine tuning and retraining them until they see the results they are looking for. While well trained models always keep the same test and train data sets separate, over time training sets even in offline systems are updated with more representative, rich and predictive data in order to build more accurate/effective models.

Clarification of responsibilities

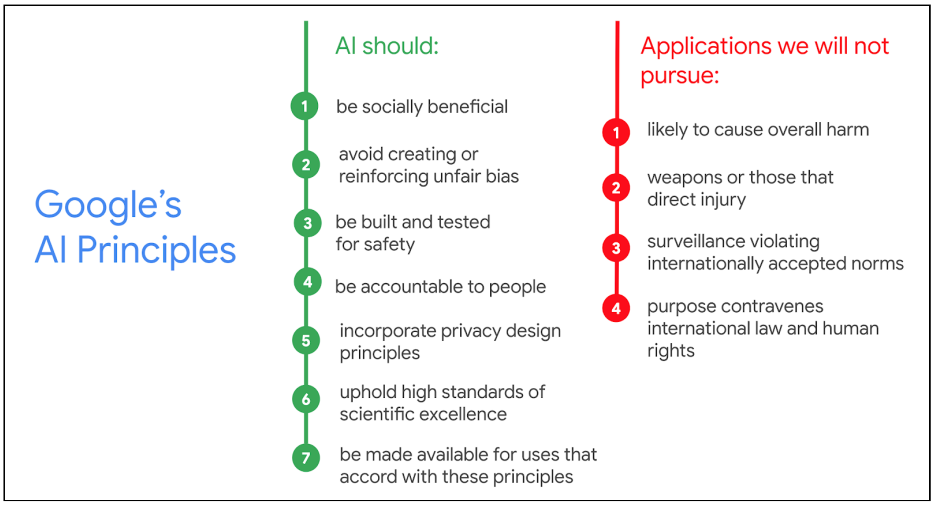
Finally, it is important to clarify the expectations for how the risk assessment for a specific AI application will be made. Just as for other laws (e.g., for GDPR in deciding if the scientific research exemption applies), Google recommends that the upfront assessment of whether a product or service is high risk is best made by those deploying it, since only they will know the intended context of its use. In particular, providers of off-the-shelf AI systems (which by nature are multipurpose) are in no position to judge, because they cannot verify the end-uses to which their systems are put.

It would be helpful for there to be an advisory body that companies could consult in confidence prior to launch, if ambiguities were to arise in interpreting the definition. In the long-term we believe sectoral regulators are the best-placed to play this role. Having consistency in oversight and the expectations for human and machine actors performing the same task would help to reduce the risk of artificial protectionist constraints being imposed, unless there are justifiable grounds for difference. However, we understand that some sector regulators may struggle if they lack the necessary AI expertise. We thus recommend a short term fix, during which a specific entity within each of the "high risk" sectors is nominated to play a lead advisory role for national regulators across Europe. This could be a particular national sectoral regulator who has already developed expertise in AI applications, or a more general European-wide sectoral body where that exists. This would help to ensure more rapid sharing of learnings, as well as making it easier for industry to collaborate in developing standards by providing a focal point for engagement.

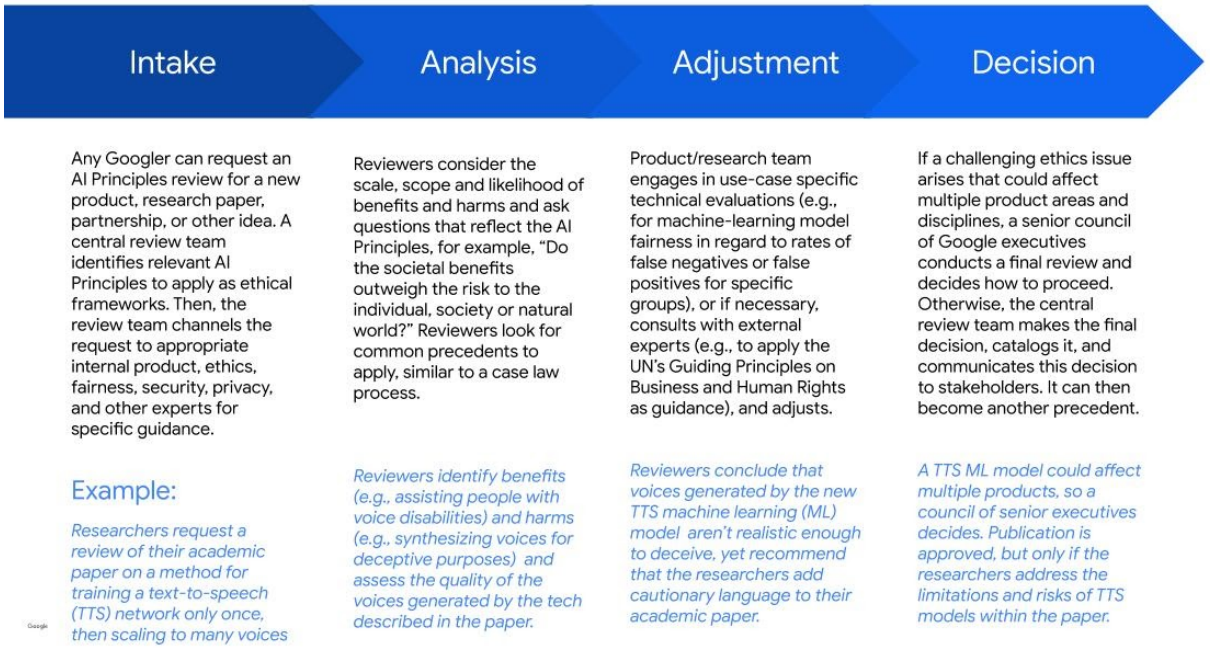
Post launch, if concerns arose that an application had been mis-classified as not high risk, remedial action could be taken via existing legal channels.

Google's AI Principles in practice

Providing guidance on how best to address ethical questions regarding the use of AI was the impetus behind Google's AI Principles. These set out Google's commitment to developing technology responsibly and establish specific application areas where Google will not design or deploy AI. They are concrete requirements which actively govern Google's research and product development and underpin business decisions:



While ultimate responsibility for living up to the AI Principles spans the company, implementation is underpinned by an internal review process supported by dedicated teams:



Rationale for ranking of concerns highlighted in the consultation survey

The consultation survey lists six concerns about AI, asking respondents to rate their importance on a scale of 1 to 5. This section provides additional commentary on the rationale behind Google's ratings, as well as providing some thoughts for further consideration.

In your opinion, how important are the following concerns about AI
(Choose from 1-5: 1 is not important at all, 2 not important, 3 neutral, 4 important, 5 is very important OR No opinion)

AI may endanger safety	4
AI may breach fundamental rights (such as human dignity, privacy, data protection, freedom of expression, workers' rights etc.)	4
The use of AI may lead to discriminatory outcomes	5
AI may take actions for which the rationale cannot be explained	3
AI may make it more difficult for persons having suffered harm to obtain compensation	2
AI is not always accurate	3

AI may endanger safety	Rated 4 = important

AI in itself is not good or bad, safe or unsafe - it comes down to how people use it. Taking a technology neutral approach in any future regulatory framework is therefore important.

While it is essential to take precautions against both accidental and deliberate misuse of AI, this must be done in proportion to the damage that could ensue and the viability of the preventative steps proposed, across technical, legal, economic and cultural dimensions. However companies and developers who are at the frontline of defence from bad actors must think carefully about the consequences of any problems their AI system could face and update their systems accordingly. This is the case regardless of root cause — be it due to predictable system failure or unpredictable behaviours, 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.

When considering safety, it is important to consider what safety precautions would be taken when using a non-AI tool and apply similar thinking to AI tools. As with non-AI tools, the appropriate performance thresholds may vary by context - in some situations it may be deemed acceptable for minimal errors, but in others such a compromise would be ethically unacceptable.

One issue that Google has long grappled with is the balance between open publication and collaboration to accelerate access and progress, and thoughtful limitations and restrictions to minimise harm. AI is a tool that can be applied with good or ill intent, and even well-meaning uses can turn out to be misguided in their real-world impact. As the ecosystem continues to evolve, it is vital that Google and others continue to evaluate inherent tradeoffs in specific innovations between the benefits of openness and the risk of abuse.

AI may breach fundamental rights	Rated 4 = important

While advances in technology have and continue to benefit people, the same advances can also restrict such rights, including freedom of expression, access to information for people of all ages, privacy, safety, and equality. In addition to recognising European law such as the Charter of Fundamental Rights, Google is committed to respecting the rights enumerated in the Universal Declaration of Human Rights and its implementing treaties, as well as upholding the standards established in the United Nations Guiding Principles on Business and Human Rights.

One of the principal ways Google ensures that human rights are protected in the use of AI is through the AI Principles, described above. These represent a steadfast commitment that Google's use of AI will avoid unfair bias, rigorously review for safety, design with privacy top-of-mind, and be accountable to people. They also specify that Google will not design or deploy AI in contexts where the purpose contravenes international law and human rights.

The use of AI may lead to discriminatory outcomes	Rated 5 = very important

AI systems are enabling new experiences and abilities for people around the globe. Beyond recommending books and television shows, AI systems can be used for more critical tasks, such as predicting the presence and severity of a medical condition, matching people to jobs and partners, or identifying if a person is crossing the street. Such computerised assistive or decision-making systems have the potential to be fairer and more inclusive at a broader scale than decision-making processes based on ad hoc rules or human judgements. But, the risk is that any unfairness in such systems can have a wide scale impact and it is critical that we work towards systems that are fair and inclusive for all.

Ensuring fairness can be a difficult task because AI models learn from data collected from the real world, and so an accurate model may learn or even amplify problematic pre-existing biases in the data based on race, gender, religion or other characteristics. Even with the most rigorous and cross-functional training and testing, it is a challenge to ensure that a system is fair in all situations. In addition, defining unfair bias is not always simple, and notions of fairness differ across cultures and societies. Fairness is often multidimensional, and optimising for one measure of fairness may require trading off another.

At Google, this is an active area of research, from fostering an inclusive workforce that embodies critical and diverse knowledge, to assessing training datasets for potential sources of bias, to training models to remove or correct problematic biases. ML fairness is an emerging area of AI research that Google is heavily invested in, and Google has launched a number of relevant open-source tools, including a What-If Tool that empowers developers to visualise biases, Fairness Indicators that help Cloud users check ML model performance against defined fairness metrics, and an ML Fairness Gym for building model simulations that explore the potential long-run impacts of ML-based decision systems in social environments.

AI may take actions for which the rationale cannot be explained	Rated 3 = neutral

AI's greatest value is seeing patterns in complex situations that are beyond human comprehension — thus (by definition) such AI systems will not be fully explainable in a way that a person can grasp. Even if the source code were shared in such a situation (an extreme form of algorithmic transparency which Google does not support) it would not help, as it would still be too complex to fathom even for experts. However, it is a fallacy that AI systems are black boxes. With enough effort and the right tools, it is possible to get some insight into why any AI system behaves in a certain way.

The problem is that explainability is costly, either in terms of technical resources or in terms of trade offs with other goals like model accuracy (if more accurate but harder-to-explain techniques have to be foregone). Tailoring explanations to be meaningful to a range of audiences is also difficult and time intensive. While there has been much progress in tools to support developers, such as Google's recently launched Explainable AI tool for Cloud AI customers, providing explanations at scale remains a challenge because the detail of what is needed varies significantly from sector to sector and across audiences.

Fortunately, just as not everyone needs to be an expert mechanic to get a driving licence and trust that a car is safe to drive, nor are explanations always necessary when using AI systems. In considering the level of explainability demanded in a specific instance, it is worth comparing the standards applied to current (non-AI) approaches. For example, an oncologist may struggle to explain the intuition that leads them 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 a reference, even if it remains a struggle to fully grasp how the data are processed to predict an 80% chance of cancer. There is a risk that innovative uses of AI could be inadvertently precluded by demanding that AI systems meet a "gold standard" of explainability that far exceeds that required of established non-AI (including human-based) approaches. 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.

Finally, it's important to acknowledge that explainability is seldom an end in itself, but rather a means of providing accountability and boosting trust. If it is not feasible to provide the desired level of explainability in a given instance, the same goals could be achieved by placing strict guardrails on an AI system's use — e.g., rigorous ongoing testing, or triggering human review if the probability of accuracy falls below a certain threshold, using interfaces that allow meaningful consideration of an AI system's output while mitigating the risk of confirmation bias.

AI may make it more difficult for persons having suffered harm to obtain compensation	Rated 2 = not important

The reason for rating this as 'not important' is because Google does not agree with the premise of the question. There are already many products and services in the marketplace with complex supply chains making responsibilities for harm opaque, for which legal frameworks exist to address. There is nothing inherently different about AI that would cause this to be more difficult than for any complex software system. So long as the notion of legal personhood for AI is avoided, every AI product or service will be associated with a person or business entity who can be held liable.

AI is not always accurate	Rated 3 = neutral

Like all systems and processes, including those that are human-based, AI will never be perfect or completely accurate all of the time. Google therefore does not believe this to be a concern. Rather the challenge is how to foster good safety practices and provide assurance that systems are reliable and secure so that companies and society can feel confident in their use.

There is a risk that innovative uses of AI could be precluded by demanding standards of accuracy for AI systems far exceeding that required of non-AI approaches. While it is important to seek to minimise mistakes, no system, whether human or AI powered, will ever be perfect, and in some situations a lower level of accuracy may be acceptable. One example is a situation requiring an urgent and immediate response, where 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[7]). A sensible regulatory standard would be to require only that AI systems perform at least to a similar accuracy standard as would be expected of a qualified person carrying out a similar task, unless there is pre-agreed justification for an exception.

Rationale for ratings and feedback on mandatory requirements

AI technologies will allow us to make rapid advances in safety, efficacy and productivity throughout society and the economy. Accordingly, any regulatory framework governing its use must be flexible in nature, not rigid or overly prescriptive - ensuring that it can accomodate rather than discourage future innovation. A challenge for regulators will be to develop a framework that is sufficiently flexible to account for this inevitable change without being so vague and overbroad so as to inject unnecessary uncertainty.

The White Paper outlines suggested mandatory legal requirements, a number of which Google is concerned could significantly hamper the development and diffusion of beneficial AI applications. Suggested ratings for each mandatory requirement can be found below as well as further feedback on why each area is important and specific feedback on the suggested requirements outlined in the White Paper.

In your opinion, how important are the following mandatory requirements of a possible future regulatory framework for AI (as section 5.D of the WhitePaper)
(Choose from 1-5: 1 is not important at all, 2 not important, 3 neutral, 4 important, 5 is very important OR No opinion)

The quality of training data sets	3
The keeping of records and data	3
Information on the purpose and the nature of AI systems	5
Robustness and accuracy of AI systems	3
Human oversight	4
Clear liability and safety rules	3

[7] One example is the use of an AI screening tool for diabetic retinopathy, which is a leading cause of blindness that is preventable if caught early. More details here: https://verily.com/stories/launching-a-powerful-new-screening-tool-for-diabetic-eye-disease-in-india/

Quality of training data sets	Rated 3 = neutral

While the quality of training data sets is an important factor, more important is how the data is used. Even poor quality ingredients in the hands of a master chef can be turned into an acceptable dish; the same is true of AI models provided that they are designed with the appropriate caveats and caution. Naturally occurring data is never unbiased, which is why Google is working on ways to meet fairness constraints even with biased training data (example).

Google's feedback on some of the specific proposals in the White Paper is as follows:

- *AI systems should be trained on data sets that are sufficiently broad and cover all relevant scenarios needed to avoid dangerous situations.*
    - The concept "all relevant scenarios" is very broad and needs to be clarified in terms of scope - the Commission should consider narrowing this to cover relevant scenarios for specific intended uses. It would also be useful to make clear what is considered a "dangerous situation".
    - Cloud AI providers are usually just providing a tool or component of a service to the customer, and will seldom know the end uses for the customer's application (akin to how the supplier of a brick will not often have insight into the ultimate construction). In such circumstances, it will be important to clarify what counts as reasonable effort on behalf of the cloud service provider to satisfy safety considerations.
    - The content of some datasets may be covered by copyright laws. For this requirement to be workable an exception to copyright for training data or special licensing agreements may be required.
- *Obligations to use data sets that are sufficiently representative, especially to ensure that all relevant dimensions of gender, ethnicity and other possible grounds of prohibited discrimination are appropriately reflected in those data sets.*
    - In Google's view, this requirement is at odds with GDPR and poses risks to users' privacy. Under current GDPR requirements, developers should not be able to access attributes such as ethnicity and therefore could not test for ethic representation in a dataset.
    - The concept of "sufficiently representative" should be better defined with respect to the range of the relevant dimensions to consider. Having a universal metric will be challenging, however, because the appropriate data quality and diversity measurements will vary by application.
    - Many AI systems are trained using multiple datasets. It will be important to clarify that the extent of representation needs to be evaluated not at the level of an individual dataset, but across the combined data corpus.

Keeping of records and data	Rated 3 = neutral

Documenting the processes followed for development and training is important, and Google encourages a strong emphasis within its teams on transparency of an AI system's performance during validation. For example, providing information about how well it performs for evaluation datasets against key metrics; providing an indication of the frequency and cost weighting that were assigned to different kinds of errors (e.g. false negatives/false positives); and if relevant, how it compares to existing human-performance benchmarks are all important.

However, it is vital that any legislative requirements for record keeping and disclosure remain sufficiently flexible to account for a wide variety of context and delivery formats. Should the required documentation be too expansive it could undermine privacy or trade secrets, or increase the risk that bad actors can manipulate the system. Google strongly cautions against making it mandatory to share the precise data used, or to reveal full details about AI models or the underlying code, as that could risk undermining business confidentiality and enable adversarial gaming of the system.

The White Paper recommends that documentation on the programming and training methodologies, processes and techniques used to build, test and validate AI systems be maintained, but much clarification is needed as to what is envisaged. In particular:

- Practical guidance will be required on how to document an AI system that has multiple algorithms that build on or feed into each other.
- Consideration should also be given as to how deep the information provided should go; in many cases, organisations will not create AI applications solely in-house but rather assemble them from components supplied by third parties or from open-source libraries.
- The most workable approach would be for the organisation providing the AI application to be solely responsible for any disclosure and documentation requirements, and third parties supplying multi-purpose AI components should be required to ensure that the terms and conditions of sale do not prevent meeting such obligations.

There are also serious practical issues relating to the White Paper proposal to require sharing of data sets in certain circumstances. For instance:

- Retaining data sets may be in conflict with some copyright provisions, particularly if non-infringement is based on only temporary use of copies.
- Providing third party access to underlying data could conflict with EU privacy laws requiring deletion of personal data, and could conflict with contractual obligations to not retain data supplied by business clients for training models.
- For products using on-device processing - so called federated learning - there is purposefully no central log of data. If an obligation to share data sets for such applications was imposed, it would undermine the significant privacy benefits of this innovative technique.
- Organisations who have built products using open-source models have no way to know the provenance of the data used to train the models unless the publisher has chosen to release it, which they will have no obligation to do (especially if they are outside of Europe).

Information on the purpose and nature of AI systems	Rated 5 = very important

As a general principle, if AI is playing a substantive role in decision-making within an application in a high risk context, that fact should be easily discoverable along with some insight into the nature of the role AI is playing, by those who have a legitimate interest. This is particularly important in cases where people could have reasonably assumed that AI was not playing a significant part — such as if AI is a behind-the-scenes tool added to enhance an existing product or service. However, it is important not to take this to an unhelpful extreme, and instead pursue a commonsense approach in considering which instances require disclosure and in what form.

Such public disclosure will typically be most appropriate for applications designed for consumer use, or which make decisions affecting individual citizens (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 optimising the operations or a wind farm or port) should not be required to be made public except in rare instances where there is deemed to be a clear public interest.

In terms of the specifics of what information should be disclosed, Google agrees with the spirit of the proposals in the White Paper, with the following clarifications and qualifications:

- *(C)lear information to be provided as to the AI system's capabilities and limitations, in particular the purpose for which the systems are intended, the conditions under which they can be expected to function as intended and the expected level of accuracy*
    - In broad lay terms, an indication should be provided as to the general logic and assumptions that underpin an AI model. It is also good practice to mention the inputs that are typically the most significant influences on output, as well as any inputs likely to be deemed sensitive or unexpected. If relevant, it is also helpful to mention 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). This is why Google is investing in scaling frameworks like Model Cards, similar in concept to widely used nutrition labels in the food industry, to increase transparency and understanding around the proper use and limitations of our AI models.
    - it is not possible to anticipate every possible use of an AI system and all possible consequences. However, it should always be possible to provide some indication 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), and broad guidance as to its appropriate use. This should be in non-technical, 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.
    - 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.

    - In practice, it will often be difficult to describe precisely the limitations and level of accuracy to be expected under different conditions in lay terms. However, research[8] has shown it is helpful for an AI application's performance to be contextualised by presenting it alongside existing human performance statistics, as well as giving concrete examples of successful and unsuccessful use cases, particularly any challenging edge-cases for humans or known pitfalls which the system has been explicitly designed to overcome.

    - It's important to retain flexibility in the format and precise details provided, because what is most appropriate will vary by context. For example, in a narrow set of domains (e.g., medicine) where expert trust heavily depends on knowing whose decisions provided the groundtruth, an indication as to the AI system's source of "groundtruth" 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.

- *(C)itizens should be clearly informed when they are interacting with an AI system and not a human being …(although) no such information needs to be provided, for instance, in situations where it is immediately obvious to citizens that they are interacting with AI systems. It is furthermore important that the information provided is objective, concise and easily understandable. The manner in which the information is to be provided should be tailored to the particular context.*

    - It should be clarified that this does not apply to AI systems that are integrated as optimisation techniques (e.g. to improve optics or another sensor within a device) on the basis that systems are not directly interacting with humans.

Robustness and accuracy of AI systems	Rated 3 = neutral

Google recommends that organisations responsible for high risk AI applications adopt a "safety by design" framework. This is a cornerstone of Google's approach to AI, with Google's AI principles requiring that AI applications be built and tested for safety. Specifically Google has committed that: *"We will continue to develop and apply strong safety and security practices to avoid unintended results that create risks of harm. We will design our AI systems to be appropriately cautious, and seek to develop them in accordance with best practices in AI safety research. In appropriate cases, we will test AI technologies in constrained environments and monitor their operation after deployment"*.

[8] "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

In terms of what this looks like in practice, Google recommends against overly-prescriptive regulation, either on a general or sector-specific level. While guidance on software engineering best practices put forward by standards bodies (e.g., the ISO and IEEE) may be a useful reference point, this should be done with caution as some established techniques for traditional software, such as formal verification (i.e., using mathematics to prove with 100% certainty that a software program meets its specification) will seldom be feasible for AI systems.

Although the general need for quality assurance and monitoring applies across all high risk AI applications, there will be substantial variation with specific applications, reflecting contextual differences in key risk factors to consider, likely trade-offs faced, and the optimal review and documentation processes. Mandating particular techniques legislatively may inadvertently undermine longer term safety by discouraging organisations from developing improved techniques and approaches. This is particularly the case for complex AI systems, such as those which involve interaction between multiple AI models. That said, monitoring for safety and accuracy throughout the life of an AI system is vital for the most critical applications. No system will ever be perfect, and most failures that occur will be unexpected.

Google's feedback on some of the specific proposals in the White Paper is as follows:

- *Requirements ensuring that the AI systems are robust and accurate, or at least correctly reflect their level of accuracy during all life cycle phases.*
    - The most appropriate approaches to ensure robustness will differ by context, although they will be rooted in some common concepts. The ability to withstand attack and "fail gracefully" is a key feature (as covered by later points), but more generally Google recommends that robustness be interpreted as affirmatively and intentionally designing the AI system to cope with failure and adapt to new situations. The precise details will vary by application and context, but common elements include coding in hard constraints to prohibit unexpected system behaviours outside of the range deemed safe and formal pre- and post-launch vulnerability testing processes.
    - Determining what counts as an appropriate level of accuracy will not be straightforward for multi-purpose products, because they are so dependent on the end-usage. It will also be challenging for applications in areas where it is difficult to judge whether a decision or action is accurate (e.g., even people disagree as to what constitutes hate speech).
    - The ex-ante assessment of an AI system's accuracy is determined by its performance in relation to a given test dataset, not its performance in the real world. Even if the AI system or the test dataset do not change, the world might. Thus, it is important to develop and deploy processes that assess the performance of high risk AI applications at appropriate intervals, based on real world experience. This could be in the form of traditional methods like surveys or customer feedback, or proactive approaches that involve selectively monitoring input data where doing so is tractable and legal.

- *Requirements ensuring that outcomes are reproducible.*
    - If the intention is to introduce a degree of traceability, it is important to acknowledge that the ability to trace outcomes from AI systems operating at scale on a daily basis will differ greatly from the deeper probing possible during upfront testing. If a stringent requirement, such as full traceability of every outcome of an AI system, was mandated, 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 the social and economic benefits of AI.
    - Clarification is needed as to what precisely counts as reproducibility. For example, should individual results be reproducible, or patterns of behaviour of a system? Must all outcomes be reproducible or just certain specified ones, or certain components thereof? Great care is required to avoid inadvertently imposing undue technical constraints. A literal interpretation of "reproducibility" would be tricky, if not technically impossible, for several reasons:
        - Not all systems will provide the identical output for the same input because some are designed with built-in randomisation. For example, systems which use differential privacy have random noise carefully injected in order to protect individual privacy. The goal of such systems is to protect privacy by making it impossible to reverse engineer the precise inputs used, while still delivering an output that's close enough to the accurate answer.
        - Using the same training data will not necessarily yield models with the same precise output 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 subsets of the training dataset. Often these data subsets are created by random sampling, in order to reduce the computational burden and allow for faster iteration, which speeds up the learning process. 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 will lead to the same model output.
            - More generally, the initial weights applied to different features within a neural network's architecture are chosen at random. These different starting points for a model's training can lead to models that perform slightly differently, even if they were trained on the identical dataset.

        - Privacy safeguards 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.

    - However, there are some potentially workable approaches that could support a broader notion of "predictability at scale", which Google urges the Commission to look to when translating this requirement into legislation. Examples include explicit versioning of code combined with information about which data was used for testing and training models (aka "data lineage"), the notion of archiving snapshots, and adopting a statistical notion of reproducibility which does not require exact matching.

- *Requirements ensuring that the AI systems can adequately deal with errors or inconsistencies during all life cycle phases.*

    - It is vital for any requirements to be flexible so that what is most appropriate can be determined by the context. For example, in settings where mistakes are hard to reverse and have extreme consequences, 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 it may be adequate to prioritise checking for anomalies and errors early and having established processes to remediate.

    - Clarification is needed as to when it is reasonable to assume that a product life cycle has ended, and what the requirements are for products that have been retired or superseded by upgraded variants.

- *Requirements ensuring that AI systems are resilient against both overt attacks and more subtle attempts to manipulate data or algorithms themselves, and that mitigating measures are taken in such cases.*

    - Currently, the best defenses against some types of adversarial behaviour are not yet reliable enough for use in a production environment. It is an ongoing, extremely active research area. Like any other type of vulnerable software, developers should think about whether their system is likely to come under attack and whether they can mitigate those attacks, consider the likely consequences of a successful attack, and in most cases simply not build systems where attacks are likely to succeed and have significant negative impact.

    - In general, however, attack resiliency for AI systems is less about algorithms, and more about processes. Good practice includes the following:

        - Building a rigorous threat model to understand possible attack vectors and the gradation of threats presented. For instance, AI systems in which users directly provide inputs may be more vulnerable than attacks on systems that only process metadata collected by a server

(like timestamps of activity) as they would be much harder for an attacker to intentionally modify.

- Testing the performance of AI systems in an adversarial setting. In some cases this can be done using tools such as CleverHans; it can also be extremely helpful to establish an internal red team to carry out the testing, or host a contest or bounty program encouraging third parties to adversarially test the system.

Human oversight	Rated 4 = important

Human input is central to an AI system's development. From problem and goal articulation, through to data collection and curation, and model and product design, people are the engine for the AI system's creation. Even with advanced AI systems able to design learning architectures or generate new ideas, the choice of which to pursue needs to be guided by human collaborators, not least to ensure choices fall within an organisation's 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 so as to enter (or remain) in real-world use. And of course, human users provide essential feedback to improve AI systems over time.

As the White Paper rightly notes, the appropriate degree of human oversight for high risk AI applications may vary from one case to another. Based on Google's experience, this is definitely the case. 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 during which such an appeal can be made. 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 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 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 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.

Clear liability and safety rules	Rated 3 = neutral

Google considers these to be of great importance. Further detail is provided in Section 3.

Concerns regarding conformity assessment concept

An important aspect of responsible decision-making regarding any new technology is assessing the risks it poses. However, it is vital to strike the right balance when doing so - ensuring that upfront requirements to enter the market are not so onerous that they hinder responsible and socially beneficial innovation, and block the use of AI in unanticipated crisis situations where a rapid response is essential.

What is the best way to ensure that AI is trustworthy, secure and in respect of European values and rules? (pick one)

Compliance of high-risk applications with the identified requirements should be self-assessed ex-ante (prior to putting the system on the market)
Compliance of high-risk applications should be assessed ex-ante by means of an external conformity assessment procedure
Ex-post market surveillance after the AI-enabled high-risk product or service has been put on the market and, where needed, enforcement by relevant competent authorities
A combination of ex-ante compliance and ex-post enforcement mechanisms
Other enforcement system
No opinion

Please specify any other enforcement system

A standalone scheme for AI systems would duplicate review procedures that already govern many higher risk products. The best approach for high risk AI applications not already part of such reviews is ex-ante self assessment, coupled with ex-post enforcement mechanisms where problems are suspected. Care should be taken to ensure the self-certification process is not too onerous, especially in terms of documentation requirements, so as to not discourage innovation or put an undue burden on SMEs.

Do you have any further suggestion on the assessment of compliance? (500 characters max)

In choosing an assessment regime, it is vital to be pragmatic to ensure it is not overly burdensome for application providers, and also practical for designated assessment bodies to deliver, taking into account the level of expertise (sectoral and AI specific) and resourcing required to implement in a timely fashion. It is also important to remember that there is no “one size-fits-all" approach to compliance, and thought should be given to the context in which the AI technology is operating.

Ex ante conformity assessment requirements as recommended by the White Paper strike the wrong balance. Creating a standalone assessment scheme for AI systems would risk duplicating review procedures that already govern many higher risk products (e.g., medical devices), adding needless complexity. For any AI systems not already part of such reviews, an approach of ex post enforcement if problems arise, coupled with clear guidance as to "due diligence" processes and expected performance standards that providers could self-assess against upfront would likely achieve similar results more expeditiously and efficaciously without risking unduly hindering innovation and erecting unnecessary burdens. An ex post regime could also more effectively build on existing industry practices, as many companies have already implemented ethical, legal and due diligence practices to guide the responsible and trustworthy development of AI.

If, however, the Commission decides to commit to an ex ante conformity assessment regime for AI systems, Google recommends that the following aspects be clarified:

- **Treatment of products already in market:** If it is required that existing products in market must undergo retroactive conformity assessments, it will create significant backlogs for newly established testing centres. A grandfathering clause would solve this at the outset. There are precedents for such treatment in other sectors, such as construction, whereby once building codes are updated, existing builds need not comply unless substantial renovations are made.

- **Treatment of R&D and early stage products:** Based on Google's experience, 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 testing and piloting of an AI application be allowed prior to any conformity assessment, within the bounds set by existing sectoral regulation. If such pre-assessment testing is not permitted, it may result in organisations taking an unduly precautionary stance when considering investments in new products, which could hinder innovation. This would significantly weaken Europe's position vis-a-vis global competitors.

- **Treatment of products which receive significant updates:** There may be a case for considering repeat assessment procedures for certain extremely high risk products that will undergo important changes during their lifetime (e.g., medical devices with embedded AI). However, if implemented, this must be accompanied by clear guidance about when such repeat assessment procedures are warranted. In particular, it is important to be aware of the context in which modifications may be made and the implications for their outputs and inputs. Potential determinants of whether a modification should spark a new assessment could include: whether there has been a change to a dataset (rather than a datatype from the same set); changes to the operating point of a model without modifying the algorithm (e.g., from 90% sensitivity and 90% specificity to 95% sensitivity and 85% specificity respectively); and whether the 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).

- **Requirement to retrain on European datasets:** The White Paper raises the possibility of requiring AI systems to be retrained using European data or in Europe, if developers are unable to prove that the original dataset used met European standards. Google has several important concerns regarding this concept:

    - It is very common in certain fields (e.g., computer vision) for training datasets to include third party and open source data. In such instances, the provenance of the training data is often a known unknown. Requiring that high risk products forgo use of such foundational and widely adopted data sets and the models that derive from them could lead to a serious degradation in the quality of AI systems that are subsequently released in the EU, particularly in instances where suitable "European" data sets do not exist.

    - From a technical perspective, it is over-simplistic to expect that retraining on European data is a panacea to solve model performance problems. It is just as possible to encounter significant fairness and diversity issues with models trained on data collected in Europe and compliant with European laws and ethics, as with data from elsewhere. In fact restricting the AI models to use only limited data sets, could lead to non-representativeness, discrimination and lower quality models, and could cut Europe off from socially beneficial innovations developed elsewhere.

    - If a model is found to fail in a European context, it is important that the model be fixed. But the manner in which that fix is made should not be prescribed by regulation, such as this requirement would impose. In some instances, failures may not even be due to issues related to the data; and even in instances where the data is at fault, there are often techniques for addressing such problems other than retraining on fresh data.

More generally, there are also effective and more workable alternatives to upfront conformity assessments that should be considered. For example:

- Prior to any launch, for AI applications deemed to be high-risk, organisations could be mandated to carry out and document risk assessments based on articulated principles. This would be analogous to the requirement for data protection impact assessments under GDPR.

    - In certain circumstances, a human rights impact assessment could be required to be carried out, undertaken by a credible expert. This would align with Section 21 in the United Nations Guiding Principles on Business and Human Rights, which many companies, including Google, are already committed to upholding.

    - There may be scope 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, could present a structured approach to documenting the expected impact of

foreseeable risks, and the corresponding preventive measures or reactive strategies planned if such failures were to occur.

- Rather than focusing on assessment of training data, Google would instead recommend rigorous testing of AI systems using curated (in some cases, potentially synthetic) test data that has the correct statistical properties to identify possible problems. This is already done for other safety-critical systems, and has the added advantage of avoiding the problems of giving third parties access to data as described earlier.
    - To help facilitate this, expert bodies could develop and publish benchmark datasets tailored to specific high-risk applications, and provide guidelines as to the performance standards and confidence levels that are deemed to be reasonable. Organisations could then conduct internal tests using these benchmark datasets, and the documented results could support a self-certification process.
    - To guard against gaming by unscrupulous developers overfitting models to the benchmark data, multiple variants of the dataset will need to be generated for use in testing. One possible approach is for a large central dataset to be created, to which new data-points are regularly added. When an organisation wishes to certify a specific application, they could request a dataset that would be randomly generated from the central dataset and registered to be used only for this certification.

Section 3 - Safety and liability frameworks

The report on the safety and liability implications of AI, IoT and robotics discusses a range of potential changes to existing EU rules, including the Product Liability Directive and General Product Safety Directive. Google agrees it is vital that safety and liability frameworks provide users of AI systems and applications with adequate protection. This will boost public trust and uptake, and ensure that the benefits of AI are fully realised.

However, Google has several concerns regarding the proposals highlighted within the report, and in the ongoing debate about the need to review the liability regime in light of AI. The EU's existing safety and liability framework is both effective and technology neutral, making it flexible enough to cover the new and emerging challenges that AI undoubtedly creates. Changing these foundational legal and societal frameworks should only be done in response to significant and demonstrable shortcomings with the current frameworks, and after thorough research establishing the failure of the existing contract, tort, and other laws. More detailed feedback on the range of safety and liability proposals being discussed can be found below.

Safety

The EU safety framework is concerned with health and safety, predominantly with the physical protection of consumers. The Commission's report appears to conflate this notion with concepts that fall outside the scope of product safety (e.g., cyber security, ethics, privacy and mental health). In Google's opinion the EU's focus should be exclusively on those areas where the unique properties of AI, IoT or robotics immediately create a heightened risk to the health and safety of consumers, such as physical injuries, chemical poisoning, choking, electric shock or fire. To the largest extent possible this should be done at the level of special safety regulation (e.g., Regulation (EU) 2019/2144 on type-approval requirements for motor vehicles). More detailed commentary on specific aspects follows.

New kinds of safety risks

The report suggests that traditional concepts of safety are challenged by AI, IoT and robotics in relation to the data dependency and connectivity of such products. It also moots consideration of mental health impacts as a safety issue in the robotics sphere. Overall such concerns appear overly broad and off-target, and the proposals to address them risk causing legal uncertainty.

The current product safety legislation already supports an extended concept of safety protecting against all kinds of risks arising from the product according to its use. However, which particular risks stemming from the use of artificial intelligence do you think should be further spelled out to provide more legal certainty? (pick all that apply)

Cyber risks
Personal security risks
Risks related to the loss of connectivity
Mental health risks

In your opinion, are there any further risks to be expanded on to provide more legal certainty? (500 characters max)

The focus should be solely on areas where the unique properties of AI, IoT, or robotics heighten the risk to the physical and mental integrity of consumers (see Attachment for more details). Note too that the selections of risk (relating to cyber, connectivity, mental health) are not intended to imply that Google believes additional laws are needed; our position is that existing laws are often sufficient, but that it would be useful to provide greater legal clarity as to their interpretation.

In particular:

- **Data dependency:** The Commission is concerned about the alleged risks to safety derived from "faulty data" and recommends specific requirements addressing the risks to safety of faulty data at the design stage as well as mechanisms to ensure that quality of data is maintained through the use of the AI products and systems. The focus on "faulty data" seems slightly off target for several reasons:
    - Data is a neutral concept that cannot be right or wrong in itself. Instead, the focus should be on the standards applied to the selection of training data for a specific model, in order to achieve the desired outcome in a particular use case (not the actual data itself, which may be appropriate for some use cases but not others). The Commission may wish to provide guidelines for companies to follow in determining what data is used to train models that directly impact user safety.
    - Training data is but one factor of many that determine the safe functioning of an AI system in a given situation. Ultimately, an AI system's output needs to pass the safety standards in current EU safety regulations, regardless of how the system arrived there.
- **Connectivity of products:** The report suggests that connectivity may compromise the safety of a product when there is a security risk that it can be hacked. However, product safety legislation is intended only to protect consumers from physical risks immediately caused by a product itself. Security and safety considerations should thus only converge when the security threat (i.e., the hacking) causes a direct safety risk (i.e., physical injury to a consumer, such as electric shock or exposure to fire). This happens only when: (1) a connected product whose sole purpose is safety (e.g., a

smoke detector) gets compromised by hacking; or (2) when the hacking intentionally and directly causes the safety concern (e.g., hijacking the controls of a connected car in order to cause a crash). Regulatory interventions conflating security and safety risks outside of these two situations is unnecessary, and would risk confusion as to what regulation applies when consumers experience a security issue versus a safety issue. To illustrate this with some examples:

    - Consider the Commission's recall of a smart watch for children in February 2019. The accompanying mobile application had been discovered to have serious security flaws, allowing hackers to communicate with and get a GPS location for any child wearing the watch. While this represented a clear security risk, it was not classed as a physical safety risk — instead authorities indicated the risk type as "other". This is because hackers would have needed to take multiple actions following the hack in order to pose any physical threat to a child, meaning there was insufficient causal link to the cybervulnerability. Such a stretch of product safety legislation would not be necessary today because the EU Cybersecurity Act (EU) 2019/881) came into law in June 2019, providing a clearer route for dealing with connectivity-related security issues.

    - A similar example would be if a hacker exploited a software vulnerability to break into a home's smart door lock and injured a person inside. Such trespass onto private property should not be regarded as a safety risk but rather as a security risk properly addressed under cybersecurity regulation mandating a fix or a recall of the product.

- **Mental health:** The Commission suggests explicit obligations for producers of AI humanoid robots to consider the immaterial harm their products could cause to users' mental health, in particular for vulnerable users such as the elderly. While the spirit of this is laudable, in practice such obligations would introduce much legal uncertainty. Specifically:

    - The issues as framed are more ethical questions around how to integrate new technologies and robots into western society,[9] rather than issues of mental health in medical and product safety terms.

    - As yet there is no commonly accepted definition for "humanoid robots" so using this terminology introduces a lot of legal uncertainty. In addition it risks being overly broad in scope, since there are many anthropomorphic features of robots (e.g., voice, eye-like sensors) that could be considered "humanoid" which are completely uncontroversial from a safety point of view.

    - The most concerning aspect, however, is the notion of "immaterial harm". This term is alien to the product safety framework and should not be introduced. Doing so would open up the scope of safety regulation and include use cases far beyond the original purpose of product safety legislation. This would put an unreasonable burden on many producers, most likely increasing prices to the detriment of consumers.

[9] Amanda Sharkey and Noel Sharkey, Granny and the robots: ethical issues in robot care for the elderly, Ethics and Information Technology, 14, 27–40 (2012), https://doi.org/10.1007/s10676-010-9234-6.

New process requirements to mitigate AI safety risks

To mitigate concerns about perceived characteristics of AI systems, a number of additional process obligations are mooted in the report. These include a risk assessment process whenever a product is subject to "important changes" during its lifetime, as well as specific requirements mitigating for opacity and to provide human oversight. These appear overly duplicative with the types of requirements outlined in the "ecosystem of trust" section of the white paper, and also to be rooted in unfounded concerns about AI autonomy.

Do you think that the safety legislative framework should consider new risk assessment procedures for products subject to important changes during their lifetime? (pick one)
Yes | No | No opinion

Do you have any further considerations regarding risk assessment procedures?
(500 characters max)

Carrying out a new risk assessment should only be required when there has been a significant change to the functionality of the product which is likely to materially alter its performance in testing or the safety disclosures made. Generic over-the-air updates (OTAs) such as security fixes, bug fixes, or simple improvements after placing a product on the market should not trigger a renewed risk assessment.

More precisely, our comments are as follows:

- **Definition of "important changes":** Generic over-the-air updates (OTAs), such as security fixes, bug fixes, or simple improvements should not be included in the definition. Only material product changes that alter product functionality in a way that impacts safety testing and safety disclosures, or that materially changes the risk assessment needed before a product is placed on the market should qualify as "important changes".

- **Requirements for human oversight:** It is important not to over-rely on human oversight as a solution to AI issues. Forms of oversight that are commonsensical in one setting will be harmful and undermine the core essence of an AI application in another. For instance, mandating human oversight could undermine the privacy and security of AI systems that run locally on a consumer's device. It would also negate the safety benefits of a product which had been designed to avoid safety issues relating to human error. More generally, requiring ongoing human monitoring and intervention would create safety and security concerns for systems designed to operate without human involvement, and deter the development and introduction of advanced technologies in Europe.

- **Mitigations for opacity:** Proclaiming "various degrees of opacity" in the decision making process of AI systems, the report suggests various requirements relating to the transparency and accountability of algorithms, as well as imposing obligations on algorithm developers to disclose the design parameters and metadata of datasets.

This would duplicate the requirements outlined in the "ecosystem of trust" section of the White Paper. Additional regulation of these aspects within the safety framework is superfluous and bears the risk of deviating and contradicting legislation.

- **Misconceptions about AI autonomy:** In general, many of the suggestions being made to mitigate AI safety risks appear to be rooted in unfounded concerns about AGI (artificial general intelligence) and misapprehensions about the prevalence of online learning AI (i.e., AI systems that learn and adapt in real time). In reality:
    - AGI does not yet exist, and many fundamental research challenges remain to be solved before it could. AI systems today are just a set of more efficient techniques for analysing data to solve very specific problems.
    - There are very few AI systems in operational use which learn "on the fly" from real time inputs. To the contrary, almost all AI models use offline learning, where AI models are frozen once they have been found to work as intended and do not change, with updated models introduced only after having been tested. In addition, arguably the technologies behind a product or application are irrelevant, so long as they function as intended and in line with public declarations (e.g., in marketing materials, press releases).

Responsibility for safety of complex products and systems

Google seconds the Commission's view that existing product safety legislation already takes into account the complexity of products or systems to tackle risks that may have an impact on the safety of users. No additional obligations are required. We also support the broad notion that the producer placing the product or service on the market has responsibility for its safety, thus providing consumers with a "one stop shop". However we have concerns about some of the proposed details of implementation and interpretation. Specifically:

- **Responsibility for complex software systems:** According to the Commission, manufacturers of the final product have an obligation to foresee the risks of software integrated in that product at the time of its placing on the market. Consequently, the report suggests additional obligations for manufacturers to ensure that they provide features to prevent the upload of software having an impact on safety during the lifetime of an AI product. While Google supports the primary responsibility of the manufacturer of record, we have a number of concerns with this proposal:
    - First, the report does not define what is an "AI product". Google advocates for a narrow definition encompassing only products that as a substantial part of their nature use an AI system. For example, many phones use AI to some degree but should not qualify as "AI products" neither semantically nor from a safety perspective.
    - Second, the scope of the suggested obligation is unclear. Clarity would be needed on what is an "upload of software" and when does it have "an impact on safety". Only material product changes that cause the product

functionality to alter in a way that impacts safety testing, safety disclosures, or materially changes the risk assessment performed before a product was placed on the market should be considered (as per earlier comments regarding misconceptions about AI autonomy).

- **Responsibilities regarding complex value chains:** Google agrees with the Commission that under the current product safety framework, no matter how complex the value chain is, the principal responsibility for the safety of the product remains with the producer placing the product on the market. This is a straightforward approach giving consumers a "one-stop-shop". However the statement that existing legislation imposes obligations to several economic operators following the principle of "shared responsibility" seems contradictory. There should be no option for the consumer to "shop" for different responsible entities in terms of joint responsibility. This type of system risks reducing the overall safety of AI systems because it could reduce incentives for smaller players in the value chain to behave responsibly, since they would be less likely to be targeted by plaintiffs seeking compensation if something went wrong.

Liability

Overall, Google believes that Europe's current liability framework remains fit for purpose, being both effective and technology neutral, so sweeping changes are not needed. There has been no evidence of problems sufficient to warrant altering such a fundamental underpinning of European law and running the risk of unintended consequences. A strong consensus among legal experts that the current framework is inadequate should be required to justify any contemplated changes.

In particular, any initiative to introduce "strict liability" for AI systems should be approached with great caution. Globally, strict liability frameworks are reserved for abnormally hazardous situations, as they preclude any consideration of intent or negligence. Introducing strict liability would mean that anyone involved in making or operating an AI system could be held liable for problems they had no awareness of or influence over, and lead to misplaced responsibility if the AI system was simply a conduit rather than the source of harm (such as if an operator used facial recognition technology developed and provided for public security purposes to instead carry out mass surveillance). Burdening AI system developers and operators with this legal exposure would have a significant chilling effect on innovation and competition as well as on the uptake of AI technology, one that would most likely disproportionately fall on European SMEs. Businesses and startups operating in Europe would also be hit disproportionately hard by virtue of the EU becoming the first global player to introduce such a change.

Do you think that the current EU legislative framework for liability (Product Liability Directive) should be amended to better cover the risks engendered by certain AI applications? (pick one)
Yes | No | No opinion

Do you have any further considerations regarding the question above? (500 characters max)

> Globally, strict liability frameworks are reserved for abnormally hazardous situations as they remove any consideration of intent or negligence. Expanding the scope of the PLD to software and all AI applications would mean that anyone involved in making an AI system could be held liable for problems for which they had no control. Google would strongly advise against burdening AI system developers with such legal exposure, as it would stifle innovation and competition.

Do you think that the current national liability rules should be adapted for the operation of AI to better ensure proper compensation for damage and a fair allocation of liability? (pick one)
Yes, for all AI applications | Yes, for specific AI applications | No | No opinion

Do you have any further considerations regarding the question above? (500 characters max)

> The existing liability framework is solid and technology neutral, making it flexible enough to cover the challenges arising with emerging technologies. Changing such a foundational legal and societal framework should be done thoughtfully and only in response to significant and demonstrable shortcomings with the current legislative framework. A strict liability regime for software would stall innovation in Europe, stifling economic growth.

However, the question of AI and liability has been a topic of long-standing debate, and so it is useful for the Commission to review the proposals being mooted. Based on the whitepaper and subsequent discussions, it appears that there are two broad changes under serious consideration:

1. An extension to the Product Liability Directive (PLD) to include software and possibly services (with the intent being to include AI systems). In practical terms this would have the effect of making their producers subject to a strict liability regime, and also introduce a number of other complications;

2. A new standalone AI liability regime, which would include alleviation (and possibly even reversal) of the burden of proof for plaintiffs in cases of negligence. It also could include strict liability and mandatory insurance for select high risk AI devices (e.g., fully autonomous robots) and services (e.g., autonomous management of electricity distribution).

With respect to these proposals, Google would like to make the following comments, in addition to the overarching problems with strict liability described earlier:

Comments on (1) proposal to extend the PLD

Software is generally more consistent with a service than a tangible, physical product, and typically does not pose the same type of heightened risks associated with traditional physical products. Therefore, rather than change the PLD and risk exposing a wide swathe of software to strict liability, a sensible middle ground would be to clarify when software should be treated as a quasi-product. In Google's view the only software that should be considered as such is software that is used in a manner more like a product than a service, and which has the potential to cause physical damage to persons or property. Such software will normally be subject to special regulation already. An example is software used as a medical device, for which the precedent is already set by virtue of its treatment as a quasi-product under Medical Devices Regulation.

Beyond the problem of strict liability described earlier, additional complications that would come with extending the PLD to software include:

- **Classification of cybervulnerabilities as defects:** The concept of a defect in a product does not translate into software, and it would be a misunderstanding to class cyber vulnerabilities as a defect. In the context of cyber risk, where a vulnerability is discovered or an attack is detected, patches are rapidly released to the software to mitigate the identified risk. It is also important to note that software producers do not fully control updates. For example, if handset manufacturers tailor the open source Android platform to suit their own specifications, then they will need to be responsible for its upkeep including issuing updates. While Google can provide updates for the original Android platform, supporting updates to open source variants is not something that Google can control or mandate. It is also not possible to force a user to accept software updates that are offered, despite their devices being exposed to cyber threats. Thus, applying the same rules to software and services as are applied to physical products would be unsuitable and unworkable.

- **Issues around "putting into circulation":** The report indicates that the notion of "putting into circulation" used by the PLD could be revisited to reflect that products may change and be altered over their lifetime, so as to better clarify who is liable for such changes. Doing this would represent a fundamental shift in product liability law, creating a huge additional burden on producers marketing products in Europe to continuously monitor and improve products indefinitely — effectively removing any time-limitation on strict liability. Such a drastic change would destroy the current well-functioning balance struck between business innovation and consumer protection.[10]

[10] See Astrid Seehafer and Joel Kohler: Künstliche Intelligenz: Updates für das Produkthaftungsrecht? EuZW Heft 6/2020, 213.

Comments on (2) proposal to create a standalone AI liability regime

The complexity of implementing a standalone regime which provides legal certainty without unduly discouraging innovation should not be underestimated. Even simply defining the scope for such a liability regime risks adding confusion, since it is likely to entail a different definition of "high risk" than that used to scope regulatory requirements. Other elements mooted (altering the burden of proof and AI insurance) also present practical challenges, and will require careful nuancing to be workable.

Additional commentary on each of these aspects follows:

- **Definition of "high risk" for liability:** Clarity over what is in scope will be vital to provide legal certainty for system operators, as well as to make the creation of insurance schemes viable. While having conflicting definitions for "high risk" is not ideal, it is a reasonable compromise given that the assessment of liability by nature requires a narrower, compensation-oriented framing than more general regulation. A possible approach could be to provide an exhaustive list of "high risk" settings that are defined as when AI is playing a significant role in instances where strict liability already applies (e.g., nuclear power plants, aviation), unless prior exemption has been granted. Such exemptions would need to be assessed on a case by case basis, but could be appropriate if the safeguards applicable to use of a particular AI application so substantially mitigated the risks such that the specific AI application was no longer deemed to present any exceptionally high risk. For example, an application of AI incorporated into a robot that must be shown to meet relevant safety standards prior to its use would have any particularly high risks ameliorated by adherence to the standard. To deter unreasonable claims against AI system operators, there should also be an exemption for cases where evidence shows that an accident was caused by another party or "force majeure".

- **Reversing the burden of proof:** As a general rule, under the liability framework we believe that alleged victims should continue to be required to prove what caused them harm. The burden of proving causation should be alleviated in light of the challenges of emerging digital technologies only if given the properties of the specific AI system establishing proof would create an unreasonable obstacle for the alleged victim. In making this determination, factors to take into account include the likelihood that the technology contributed to the harm (e.g., if there are known defects), the nature and scale of the harm claimed, the degree of ex post traceability of contributing processes within the technology as well as the degree of ex post accessibility and comprehensibility of data collected and generated by the technology. It should be up to the alleged victim to prove that, all things considered, the burden of establishing proof for the negligence on behalf of an operator or developer of the technology, a defect in a product or the causal link between the latter and the damage is unreasonable.

  The same section of the report also suggests that a product that does not meet mandatory safety rules could be considered defective. This would contradict established case law, which allows flexibility in how a producer compensates for any

deviation from required standards.[11] Such a direct translation would create a new class of defects by law that did not previously exist, and do nothing to alleviate the burden of proof for the victim.

- **AI insurance:** Any proposal to make insurance mandatory for AI systems would require clear buy-in from the insurance industry, which has previously raised doubts about the viability of creating one 'AI insurance' - instead suggesting that it would be necessary to develop specific schemes for particular applications. A potential alternative that should be explored is specific insurance schemes for identified high-risk applications, with a fallback blanket system for other AI applications.

  It's also important to be cognisant of costs that AI device manufacturers would incur associated with an insurance scheme, potentially even for very low-risk applications, which would likely be passed on to consumers. In structuring such a scheme it will be important to discourage companies with insurance opting to impose less stringent standards due to the safety net that insurance provides.

[End]

[11] Ibid.