# Consultation on the Al Inception Impact Assessment Google's submission - 9 September, 2020

## Introduction

Google welcomes the opportunity to provide further input to the European Commission's deliberations around Al governance. The promise of Al to deliver societal benefits cannot be realised without well-founded public trust in Al's use. If done well, establishing oversight mechanisms will reassure the public that there are controls in place, while also providing useful clarity and directional guidance for industry — both of which are vital for a thriving and innovative Al ecosystem. The challenge is to ensure that any interventions in the development and use of Al — especially those that are mandatory and impose upfront costs — are suitably tailored and appropriately balanced so as to have the desired effect with minimal unwanted repercussions.

This document is intended as a companion to <u>Google's detailed submission</u> to the earlier White Paper consultation, responding to the specific options laid out in the Commission's <u>Al inception impact assessment</u> published in July.

## **Definitions**

There are several terms in the Inception Impact Assessment for which a question of definition arises on which we offer specific comments:

• Definition of AI — In the discussion of policy option 4, it is noted that a "core question relates to the scope of the initiative, notably how AI should be defined (narrowly or broadly) (e.g. machine learning, deep neural networks, symbolic reasoning, expert systems, automated decision-making)". There is no answer to this that everyone will agree on, but there are some definitions that will be more appropriate than others. Google has already put forward a suggestion in our White Paper submission based on the High Level Expert Group definition; an alternative but similar approach that may be more resonant globally is that put forward by the OECD's expert group.

The definition of AI is crucial not just in terms of determining what is in and out of scope, but as importantly because of the impact it will have on technology development and adoption by European businesses. For instance, if AI were defined as broadly as 'automated decision making' and permitted only after undergoing onerous procedural review, it will slow the uptake of automated decision making of every stripe. In some instances this may help to eliminate overly risky and potentially harmful uses; but in most cases it risks deterring and delaying European businesses from employing safer, more

reliable, and faster Al-based approaches to decisions. This would go against the EU's stated desire to create an ecosystem of excellence for Al in Europe, as it would put European companies at a competitive disadvantage. Furthermore, how "automated decision making" is defined (e.g., what counts as a "decision"?; what role does an algorithm need to play operationally to count as "making" a decision?) and implemented (e.g., for all systems or only newly developed systems?) could also inadvertently bring many innocuous automation tools and applications into regulatory scope.

In short, it is critical to strike the right balance in defining the AI applications that are in scope, focusing on those that are more likely to present real risks now and into the future, but at the same time being careful not to inhibit EU innovation and the economic benefits of various automated systems in an ever more competitive global environment.

• Definition of immaterial harm — If such a term is to be used in legislation, it will be important to provide an exhaustive and specific catalogue of covered harms. The inception impact assessment takes a step towards this by suggesting that "immaterial harm can relate to, for instance, loss of privacy or limitations to the right of freedom of expression ... or unlawful discrimination ... or discrimination in the access to products and services such as lack of accessibility for persons with disabilities, to name but a few." This definition, however, is still too open-ended. An alternative approach if a broader framing is desired could be to avoid use of the term "immaterial harm" and instead refer to "significantly restricting the exercise of fundamental rights". This would have the advantage of clarity by anchoring the definition in a body of existing law, and would appear (based on the description here) to match directionally with the Commission's ambition.

## Policy options

The Inception Impact Assessment lays out five options for policy development. Our comments on each are as follows.

Baseline (no EU policy change, Option "O"): in the absence of EU action on establishing specific requirements for AI, the risks linked to the latter would remain unaddressed. While EU legislation on the protection of fundamental rights and consumer protection as well as on product safety and liability remains relevant and applicable to a large number of emerging AI applications, problems with enforcement of existing EU law and national liability rules may emerge. There could also be a lack of clarity regarding possible obligations to address the new risks raised by AI. Given the significant commercial opportunities offered by AI solutions and the pressure to conquer market shares, whether by European or foreign developers interested in the EU market, "untrustworthy" AI solutions could ensue, with a likely backlash against AI technology as a whole by citizens and businesses. Moreover, as indicated, no EU policy change could lead to increased fragmentation due to interventions at Member States level.

As Google's CEO said in January, we agree that AI is too important not to be regulated — the only question is how to approach it. There are already many regulations and legal codes that are technology neutral in nature, and thus broad enough to apply to AI. Clarifying expectations as to

how these existing rules should be interpreted for Al applications would be extremely helpful, both in providing guidance for industry as well as pinpointing any areas where change is needed.

While we do not dispute that current legislation may have some gaps, we disagree with the premise that risks linked to Al can only be addressed by putting in place new bespoke legislation. In some instances the problem could be simply due to a lack of awareness and poor enforcement, or lack of clarity in applicability of existing rules. If existing rules were highlighted and resources invested to support swifter and appropriately calibrated enforcement, even a baseline of no policy change could result in substantial ecosystem improvements.

For example, consider the issue of discrimination and AI. Existing EU legislation<sup>1</sup> already includes the concept of "indirect discrimination" in a manner that is broad enough to encompass inadvertent discrimination by algorithms, including AI systems. However it would be useful to provide formal commentary regarding how it should be translated, in particular in terms of what are plausible, reasonable grounds for lawful discrimination in sample contexts, and what counts as satisfactory evidence of non-discrimination in instances of AI systems that are not fully explanatory.

Similarly, guidance issued in 2019<sup>2</sup> by the Medical Device Coordination Group (MDCG) on the qualification and classification of software as a medical device or accessory is broad enough to apply to AI systems, although it could be clarified if AI changes anything relating to the respective classification as described in the guidance.

Option 1: EU "soft law" (non-legislative) approach to facilitate and spur industry-led intervention (no EU legislative instrument): Under this option, EU "soft law" would promote industry initiatives for Al. A large number of Al principles and ethical codes have already been developed by industry actors and other organisations. In the Union, the HLEG developed a set of Ethics guidelines for trustworthy Al with an assessment list aimed at providing practical guidance on how to implement each of the key requirements for Al. The "soft law" approach could build upon existing initiatives and consist of monitoring and reporting on the voluntary compliance with such initiatives based on self-reporting; encouraging industry-led coordination on a single set of Al principles; awareness raising among developers and deployers of Al systems around the existence and utility of existing initiatives; monitoring and encouraging the development of standards.

Active engagement by industry in tackling AI issues is crucial, because those who are at the frontlines of developing and applying AI are in a prime position to help spot problems and propose workable remedies. Regulation will never be perfect in covering every eventuality, so it is important that regulatory compliance is not seen as an acceptable excuse for neglecting to stay alert to wider complications. No matter what policy options are pursued, lending support to industry in establishing norms of responsible behaviour and sharing best practice learnings is worthwhile.

<sup>&</sup>lt;sup>1</sup> Specifically, the Racial Equality and the Goods and Services Directives say that discriminatory decisions that are mediated by an 'apparently neutral provision, criterion or practice' but still disproportionately affect certain protected groups are qualified as indirect discriminations. In an Al context, quite obviously, the algorithm itself constitutes an 'apparently neutral criterion'. Thus, to the extent such criterion is indeed not neutral but perpetuates bias, 'algorithmically mediated discrimination' will generally fall within the scope of indirect discrimination. For a fuller analysis of this see Philipp Hacker, Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law (Common Market Law Review 2018, 1143-1185)

<sup>&</sup>lt;sup>2</sup> https://ec.europa.eu/docsroom/documents/37581

The focus however should be on supporting initiatives that are likely to make a genuine difference to what happens in practice operationally. In this respect, there is little to be gained by having a single set of Al Principles when there is already strong alignment across those that exist. Similarly, there is little need to incentivise the development of Al standards when there is already so much momentum and progress being made — not only within the confines of established standard setting bodies like ISO, but also through more informal industry benchmarking collaborations such as MLPerf.

Based on Google's experience, there are gaps in the understanding of Al issues between individual expert practitioners whose extensive experience has fine-tuned their intuition for spotting likely problems more so than those newer to the field whose senses are not yet honed. For this reason Google has made it a priority to share the materials used to train our employees when they too were first learning to grapple with challenges posed by designing Al systems that learn from data (e.g., the 15 hour Machine Learning Crash Course, 4 hour Testing and Debugging in Machine Learning course, Fairness in Machine Learning module and more, available for free at the Learn with Google Al site. The Commission could help in packaging up, highlighting and building upon such skill-building resources available from Google and others as part of promoting a responsible Al ecosystem.

Another challenge is the often subtle gap in practical understanding between those with academic or media-based knowledge of potential issues, and those who have tackled AI issues and tradeoffs firsthand in product development. The Commission can help to bridge these differences by level-setting the expectations for 'due diligence', and by providing incentives and opportunities for respectful collaboration and cross-pollination of viewpoints. The goal should be to establish clear and concise expectations that reflect the views of a broad set of stakeholders in terms that can be easily understood and efficiently integrated into product development workflows by AI experts.

Option 2: EU legislative instrument setting up a voluntary labelling scheme: Under this option, a EU legislative instrument would establish a voluntary labelling scheme to enable customers to identify AI applications that comply with certain requirements for trustworthy AI. While participation in the labelling scheme would be voluntary, the economic operators who choose to participate would have to comply with certain EU-wide requirements (in addition to existing EU legislation) in order to be able to display a quality AI label. The label would function as an indication to the market that the labelled AI application is trustworthy. The voluntary labelling scheme could follow a model similar to, or be inspired by, the assessment list of the Ethical guidelines piloted by the HLEG. A label could also be used for issues that go beyond regulated aspects and the respect of fundamental rights.

Google remains skeptical of the value of a labelling scheme — voluntary or otherwise — as a vehicle for influencing consumer behaviour in their choice of Al application. While there may be some contexts where it helps (e.g., public sector use of Al, where it could be a requirement for procurement), in other settings it is likely to be only a minor influence, with most consumers favouring functionality over labels or following the advice of people they know, be it an expert such as their doctor or a friend or colleague.

If the Commission decides to proceed with a labelling scheme, we urge caution in the scoping and operational framework imposed, to reduce the burden on SMEs, and ensure there is flexibility to evolve in line with emerging standards.

In this respect, while the <u>Assessment List for Trustworthy AI (ALTAI)</u> developed by the High Level Expert Group (HLEG) is a useful reference and exhaustively covers a number of critical issues, a less strident approach might be more compelling in fostering uptake in key application areas. As it stands, the final version of the ALTAI is over 20 pages long and the pointed yes/no nature of the questions about actions taken implies that all efforts are appropriate for every context. In reality, there will be much variation across settings as to what is relevant and most effective. Rather than basing a labelling scheme on a long and generic checklist, a better approach would be to work with industry to craft application- and context-specific versions that prioritise the key considerations and actions that will be most impactful in each setting.

For example, rather than a binary set of yes/no questions as in the ALTAI section on "Avoidance of Unfair Bias", a more helpful prompt for application developers would be open-ended questions accompanied by performance benchmarks. See Table 1 for a (non-exhaustive) illustration of how such questions could be framed to foster a thoughtful and responsible approach. Such an approach would not only provide better directional guidance to those implementing AI systems, it would also enable richer insight for regulators and consumers of the factors considered and trade offs associated with a given product or service.

Table 1: Comparison of alternative approaches

Extract from ALTAI: questions from section on Avoidance of Unfair Bias

Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the Al system, both regarding the use of input data as well as for the algorithm design?

**Did you consider diversity and representativeness** of end-users and/or subjects in the data?

- Did you test for specific target groups or problematic use cases?
- Did you research and use publicly available technical tools, that are state-of-the-art, to improve your understanding of the data, model and performance?
- Did you assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the Al system (e.g. biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness)?
- Where relevant, did you consider diversity and representativeness of end-users and or subjects in the data?

Did you put in place educational and awareness initiatives to help Al designers and Al developers be more aware of the possible bias they can inject in designing and developing the Al system?

ALTERNATIVE: illustrative (non-exhaustive) open-ended questions and benchmarks for avoiding unfair bias

How have people in different groups been historically affected by this kind of product or use case? How will they be affected by this product? Consider the product at issue, not just how groups have been affected by such technologies generally.

Recommended: No groups have been historically negatively affected by this specific kind of product; OR one or more groups have been negatively affected by this specific product or use case in the past, but this particular product offers clear benefits to these groups and there has been significant user testing to show this product will be beneficial, and the product team has consulted with relevant internal/external experts.

Requires additional work: One or more groups has been historically negatively affected by this specific kind of product, but it is possible that this particular product offers some benefits to these groups. An evaluation is needed to assess the impact of this product on these groups, and pre-launch user testing should be conducted.

How does the product perform across different user types (e.g., gender, age, skin tone, face/body feature shapes, effect of lighting, effect of makeup/clothing, language, disabilities)? In determining what performance distribution is acceptable, consider factors including: what groups may be most impacted by false positives and

Did you ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the Al system?

- Did you establish clear steps and ways of communicating on how and to whom such issues can be raised?
- Did you identify the subjects that could potentially be (in)directly affected by the Al system, in addition to the (end-)users and/or subjects?

Is your definition of fairness commonly used and implemented in any phase of the process of setting up the Al system?

- Did you consider other definitions of fairness before choosing this one?
- Did you consult with the impacted communities about the correct definition of fairness, i.e. representatives of elderly persons or persons with disabilities?
- Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness?
- Did you establish mechanisms to ensure fairness in your Al system?

negatives, performance benchmarks of existing relevant products or previous-generation technologies, existing human levels of accuracy and using noise in human ground truth as an upper bound, expectations from user studies, published research on accepted standards and failure rates within given sectors and communities.

Recommended: There is a clearly defined distribution that would be appropriate for variant performance across groups, based on user testing, noise in human rater ground truth, published research, legal requirements, and other relevant inputs. This product has been tested across a diverse set of user groups and meets or is narrower than the target performance distribution among groups. Where appropriate, the performance distribution and/or determination process is shared to provide users with more information and the opportunity to compare with alternative products.

Requires additional work: The product performs at a significantly wider distribution than the defined target for one or more groups; OR the product has not been sufficiently tested for fairness; OR various categories of target users have not been considered or identified.

Option 3: EU legislative instrument establishing mandatory requirements for all or certain types of Al applications: Under this option, the EU legislative instrument would establish certain mandatory requirements on issues such as training data, record-keeping about datasets and algorithms, information to be provided, robustness and accuracy and human oversight.

- As a first sub-option, the EU legislative instrument could be limited to a specific category of AI applications only, notably remote biometric identification systems (e.g. facial recognition). Without prejudice to applicable EU data protection law, the requirements above could be combined with provisions on the specific circumstances and common safeguards around remote biometric identification only.
- As a second sub-option, the EU legislative instrument could be limited to "high-risk" Al applications, which in turn could be identified on the basis of two criteria as set out in the White Paper (sector and specific use/impact on rights or safety) or could be otherwise defined.
- In a third sub-option, the EU legislative act could cover all Al applications.

In general, while Google is supportive of principled constraints and expectations of due diligence to ensure responsible use of AI, it is important to avoid overly prescriptive requirements. The more stringent and tightly defined the requirements are, the less flexibility there will be to adapt them as the technology evolves. The benefits of regulation in reducing real and perceived risks must also be balanced against the opportunity cost of lost societal benefits from innovation and economic growth that can result from over-regulation relative to other markets. The goal should be a properly balanced regulatory approach that builds trust in AI, provides greater clarity for technology development and business planning, and maximises the social and economic good for Europe's citizens.

Google's thoughts on the specific mandatory requirements proposed are given at length in <u>our submission</u> to the White Paper consultation, with key concerns summarised in Table 2. Therefore we have limited the rest of our commentary in this section to just the question of scope.

Table 2: Key concerns with proposed mandatory requirements

Keeping of datasets should not be mandated: Keeping datasets is likely to conflict with GDPR provisions requiring deletion of personal data, as well as presenting challenges for copyrighted datasets authorised for only short-term access. It would destroy the privacy benefits of on-device processing because it would effectively force data to be collected and stored centrally. It would prevent the use of off-the-shelf, open-source models, since developers will generally have no access to the data used to train them.

Too much emphasis on training data quality; not enough on testing output: With enough expertise and care, it's possible to create a high-performing model even using biased, low-quality training data. Thus, rather than putting requirements on training data, it would be better to have requirements based on testing model performance using benchmark datasets, to make sure that the outputs are within an acceptable range, since it is the model output that ultimately determines the real world impact of an Al system. In addition, the proposed obligations for developers to "ensure datasets are sufficiently representative" conflict with GDPR under which developers are not meant to have access to sensitive attributes like ethnicity. It is also unclear how to determine what is "sufficient", especially for providers of multipurpose Al systems.

Avoid taking a literal approach to "reproducibility": The whitepaper proposes "requirements ensuring that outcomes are reproducible". A too literal interpretation of reproducibility would be impossible to satisfy, as many Al systems have randomness built in, which makes it impossible to guarantee you get the identical output every time even if the input is the same. To be workable, there will need to be scope for broad notions of "predictability at scale" that do not require exact matching.

Imposing mandatory requirements on only a bounded and clearly defined set of applications provides legal certainty, and enables precise targeting of measures where they are most needed. It provides a framework to which additional applications can be added over time, if evidence of concrete harms emerge. It also reduces the scale of resourcing needed for enforcement, and a narrower focus will make it easier to equip teams with the necessary expertise to carry out assessments.

- Regarding sub option 1: Given that there is already a robust discussion in civil society
  regarding the use of facial recognition systems, and it is clearly an area of higher risk,
  remote biometric identification systems seem a good example of an application to which
  mandatory requirements and safeguards could be usefully applied.
- Regarding sub option 2: Conceptually, Google supports a risk-based approach to a new regulatory framework, but believes it is important to ensure that any potential regulation is targeted at the right use cases, taking into account the likelihood of harm and not just the severity of the harm, as well as a nuanced consideration of the opportunity cost (that is, the forgone benefits) of not using Al. We outlined a number of factors to take into account in our earlier submission. However, in terms of process it may help to frame risk assessment as having several distinct steps in questioning:
  - Q1. What is the inherent risk of applying this technology to this specific problem? While there may be some cases where a certain technology is in itself inherently risky, more often the primary driver of risk will be derived from the use context.
    - E.g., even in a sensitive field such as law enforcement, the risks of facial recognition technology will vary by application. For instance, using facial recognition to authenticate

officers' identities carries different risks than using facial recognition to conduct surveillance and identify criminal suspects.

- Q2. How do the attributes of this particular AI system impact overall risk?

  Specific design features and operational constraints both technological and in terms of business processes may justify reducing or increasing the overall assessment of risk from its inherent level.
  - E.g., in some use contexts, a system that has been designed to perform reliably with a similar degree of accuracy across demographic groups may be deemed less risky than a system that is often more accurate but has less consistent performance across groups.
  - E.g., in some use contexts, a system deployed to operate autonomously may be deemed more risky than systems with established monitoring procedures and other operational checks in place.
- Q3. What additional mitigations can be put in place to better manage overall risk?

  A wide variety of mitigations should be considered, including technical mitigations (e.g., utilising additional data points to improve accuracy), operational mitigations (e.g., minimum confidence thresholds), and environmental mitigations (e.g., physical barriers between autonomous systems and human beings).
- Q4. What is the overall risk of this application, and is this tolerable when compared to existing alternatives?
  - Al systems will never be perfect but nor is human decision making. Taking into account the inherent risks (Q1) of the technology and use case, specific attributes of the particular system (Q2), and any additional mitigations that can be put in place to manage those risks (Q3), how risky is the Al system overall? Is this remaining risk greater or less than the risks of not using Al, and is it tolerable?
    - E.g., if the overall risk of an Al system is deemed less than the risk inherent in established methods of carrying out a task, arguably it should be tolerable.
    - E.g., in some use contexts, it may be tolerable to accept an Al system that has higher risk
      and performs more poorly than a human expert carrying out the same task, if there are not
      enough people willing and able to do the job.
- Regarding sub-option 3: Making no distinction between higher and lower risk AI applications makes it far harder to create regulation that is proportionate and effective, going against the EU's better regulation principles. If mandatory requirements are set to a level appropriate for higher risk applications, they will be unnecessarily burdensome for those that are lower risk (likely the vast preponderance of applications) creating additional costs, delay and friction for European developers of all sizes. This would significantly hamper Europe's ambition to increase the uptake of AI, and needlessly disincentivise innovation that is lower risk, thus slowing or even preventing development of many socially beneficial applications. At the other extreme, if mandatory requirements are set at a level to avoid imposing overly burdensome requirements on low risk applications, they may not be sufficient for those that are of higher risk.

In addition, the cost of enforcing all-encompassing regulation would be far higher for regulators, in terms of building the necessary operational infrastructure and equipping an army of people with skills spanning every possible application context required to assess and regulate <u>all</u> Al applications. The time it would take, and scale of investment, training

and coordination needed to support blanket regulation across all Al applications should not be underestimated.

Option 4: combination of any of the options above taking into account the different levels of risk that could be generated by a particular Al application: Any of the sub options above can be combined with industry-led intervention or not. In one scenario (industry-led intervention or co-regulation), the legislative instrument would primarily consist in high-level principles and obligations to be complemented by industry-led norms such as in the form of standards or codes of conduct. In another scenario (detailed regulatory framework), the legislative instrument would establish a regulatory framework with a higher degree of detail and specificity, possibly to be achieved also with implementing powers of the Commission.

Allowing a combination of approaches (ranging from the baseline of option 0 through to option 3) seems a commonsense way to ensure proportionality, offering the greatest flexibility to tailor how potential harms are addressed without making overly expansive and burdensome demands on applications that pose little risk. This approach would also provide a sensible foundation for evolution in light of practical learnings that arise and technology developments, and provide due consideration for real-world resourcing limitations.

## **Enforcement mechanisms**

The Inception Impact Assessment notes that for any of the policy options:

"...the EU legislative instrument should include enforcement mechanisms to ensure effective compliance with any applicable requirements, in particular existing and future requirements under the EU acquis. Such mechanisms could be ex-ante or/and ex-post. Ex-ante mechanisms could consist of conformity/safety assessment procedures that are aligned with the procedures that already exist in the product safety legislation. For Al applications where no mechanisms exist, new ex-ante conformity/safety assessment procedures may need to be established. Moreover, existing ex-ante and ex-post enforcement structures would need to be competent and fully equipped to fulfil their mandate where Al tools are used. Ensuring this competence may include requirements for adequate funding, capacities, competences and mechanisms to work together"

Google welcomes the acknowledgment of the importance of aligning any ex-ante assessment with existing procedures, and ensuring that enforcement structures are suitably resourced and competent to carry out their role. The scoping of any new rules (or heightened enforcement of existing ones) will determine what is necessary, and so it is vital that a pragmatic approach is taken. Any new rules should take into account the implementation challenges for both regulators and those regulated alike.

Just as there is a need for proportionality in the nature of requirements, so too is there a need for similar balancing in enforcement mechanisms, including leveraging self-assessment procedures where appropriate. Except in fields where upfront third party reviews are already established practice, Google recommends 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.

For example, if there were differing requirements depending on the risk classification for an application, there needs to be an upfront decision made as to what is the risk classification for

any given application. A practical approach would be for regulators to provide detailed templates and guidance on how to carry out and document the risk assessment, but delegate responsibility to those using the AI system to conduct it. Whether carried out by internal expert teams with a variety of backgrounds (technical, ethical, legal etc) or by enlisting trusted third party consultants, the resultant documentation would provide evidence of its satisfactory completion, and could be available to view on demand by regulators, or even be filed confidentially with a certification body if more routine transparency is desired.

While an application-specific assessment will make sense for individual high risk applications, if regulation is scoped more broadly, a more scalable approach may be to focus instead on certifying a company's internal governance processes rather than every single use of Al in their operations. For example, companies could conduct periodic audits of their governance processes using independent expert auditors who are professionally qualified and entrusted to only certify organisations who meet the appropriate standards.

#### \*\*\*\*\*

Google welcomes the Commission's continued outreach and momentum in establishing an Al governance framework in Europe. Without a foundation of trust, the opportunities that Al offers will not be fully realised. We remain committed to engaging constructively and sharing learnings from our internal governance process, which is central to ensuring our responsible use of Al and instilling trust in our services. We look forward to providing further comments as the Commission's thinking develops.