JUNE 14 2020

CHARTING A WAY FORWARD

FACEBOOK's Comments on European Commission White Paper on Artificial Intelligence - A European Approach

FACEBOOK

Table of Contents

Introduction and Summary 01

As a company that supports proactive regulation around novel **technology policy issues**,[1] Facebook welcomes the European Commission's White Paper on "Artificial Intelligence - A European approach to excellence and trust" (the "**White Paper**"),[2] and is eager to collaborate on the future of AI governance. We agree with the Commission that AI is a uniquely powerful technology that must be developed, deployed and used responsibly.

Facebook, whose wide range of AI-enabled products and services are available throughout the European Union, supports the Commission's goal of investing in an "ecosystem of excellence" and building an "ecosystem of trust" around AI in Europe. We agree that the EU, as a global leader in digital technology with a growing and thriving AI industry, can and should leverage AI to boost its research and industrial capacity, while increasing its competitiveness and strengthening its economy. Given the EU's foundational values of respect for human rights, democracy, and the rule of law, combined with its leadership role in advancing a human-centric vision for Trustworthy AI, the EU is also in a unique position to establish sensible standards around how to maximise the benefits of AI while responsibly managing its risks. In showing such leadership, it will encourage the development of a transatlantic AI governance framework – and, ultimately and hopefully, a global standard – rooted in shared respect for fundamental rights and democratic values.

There is a fast-growing global dialogue around how best to turn broad responsible AI principles into practical steps that both companies and policymakers can implement, thanks in great part to the efforts of the Commission itself. That is why we broadly recommend in our comments that any new AI regulation should support and build on these ongoing efforts to establish best practices, rather than risk cutting them short with inflexible rules that may not be able to adapt to a rapidly-changing field of technology. More specifically, our comments focus on two primary recommendations:

Clearly Defining High-Risk AI.

Facebook is aligned with the Commission's goal of limiting regulation to those highest risk AI uses that require it – the question is how to define high-risk AI. Facebook urges the Commission to be precise in defining "AI" and what sectors and subsectors it considers high-risk, and when defining what counts as high-risk AI applications within those sectors, we urge the Commission to avoid broad undefined terms or exceptions like "immaterial damage" or "exceptional circumstances."

Aligning with GDPR Around Self-Assessment of AI Risk.

We generally urge that any new AI regulation should build upon the requirements that already exist in GDPR, to provide greater legal clarity, avoid duplicative regulation, and ensure a proportionate approach to these novel issues. In particular, we highlight how the Commission's proposed system of enforcement – requiring prior conformity assessments of AI systems by regulators or third-party auditors before those systems are deployed in the EU – risks unnecessarily overburdening AI developers and significantly impairing innovation and economic growth that would benefit European citizens. As a more balanced alternative structure, we point to GDPR's approach based on companies' self-assessment of risk, with regulatory enforcement when companies fail to properly conduct a risk assessment or mitigate the risks they identify. An “Automated Decision-making Impact Assessment” approach similar to GDPR’s “Data Protection Impact Assessments” could help accomplish the Commission’s goals in a much more proportionate and flexible manner.

We also raise, in the final section of our comments, a range of practical questions and concerns around specific elements of the White Paper’s proposal, including highlighting tensions between the proposal and other legal obligations like those around data protection and intellectual property, and raising technical concerns around some of the proposed mandatory requirements based on our practical knowledge and experience as an AI company.

As our comments highlight, AI poses complex new challenges to existing legal frameworks, and deciding what an effective and technically feasible AI regulation should look like will not be easy. But we are eager to continue the conversation and collaborate with the European Commission and other policymakers on these hard questions, not only in the EU but around the world.

02 FACEBOOK'S Work in Support of Excellence and Trust in AI

AI is the core technology that enables Facebook to achieve its mission of giving people the power to build community and bring the world closer together. Our large team of researchers and engineers, based in Europe and globally, develops machine learning algorithms that rank feeds, ads and search results in Facebook and Instagram, and creates new text understanding algorithms that help to keep spam, misleading content, and other violations of our **content policies at bay**.[3] New computer vision algorithms can "read" images and videos to the blind and display over 2 billion translated stories every day, while new speech recognition systems can automatically caption the videos our users post. AI can make our existing products better while also enabling entirely new experiences, and it is absolutely instrumental in keeping people safe on our platforms.

As an AI-driven company, Facebook stands ready to constructively participate in the effort to foster both excellence and trust in the AI ecosystem, a goal that has driven much of our work in recent years.

A. FOSTERING EXCELLENCE IN AI

Facebook is pleased to see the Commission prioritising the development of the AI technology sector in the EU, with a starting focus on enhancing research and innovation while incentivising the adoption of AI-based solutions in business and government. Below are just a few examples of how Facebook itself has been actively working to advance the state of the art in AI research and to leverage AI for positive impacts, contributing to an AI "ecosystem of excellence" in the EU and around the world.

Offering Open Source AI Research and Tools.
Facebook doesn't just invest in AI for its own products but for the world, supporting the development of cutting-edge AI research and state-of-the-art open source tools through Facebook AI Research (**FAIR**).[4] Founded by computer scientist and scholar Yann LeCun, FAIR's international network of collaborating researchers is invested in long-term and foundational research to unlock the full future potential of AI, and is dedicated to sharing that research openly. FAIR has multiple hubs around the world including **FAIR Paris**,[5] a world class research lab that is celebrating its 5th anniversary this month with an online interactive conference highlighting its achievements over the past half-decade and unveiling a range of new European AI research projects. Indeed, FAIR-affiliated researchers from a broad international selection of academic institutions, including in the EU, publish a steady stream of innovative research papers – often accompanied by supporting data sets and models – that broadly cover the full range of AI theory, algorithms, applications, software infrastructure, and hardware infrastructure across deep learning, computer vision, natural language processing, speech, and reasoning. FAIR and the broader Facebook AI team also publishes and maintains a broad range of open source AI tools including PyTorch, a fast and flexible deep learning framework used by countless developers and researchers across the EU and around **the globe**.[6]

Supporting Innovative Academic Initiatives.
Facebook is dedicated to supporting innovative academic research into issues of AI ethics and governance, including in the EU. Most notably, Facebook has partnered with the Technical University of Munich (TUM) to support the creation of an independent **Institute for Ethics in AI (IEAI)**.[7] Drawing on expertise across academia and industry, TUM IEAI conducts independent, evidence-based research on fundamental issues that affect the use and impact of AI, such as safety, privacy, fairness, and transparency. To help support thoughtful and groundbreaking academic research that takes into account other regional perspectives, we also launched a call for papers in India and the Asia-Pacific region on topics regarding AI governance, cultural diversity, and **ethics by design**,[8] and have joined forces with

CETYS and the Inter-American Development Bank (IDB) to launch GuIA.ai., an AI academic research project focused on Latin American and Caribbean perspectives on **AI governance and ethics**.[9]

Driving Independent Research Through "Innovation Challenges."

To incentivise outside research on hard AI problems, Facebook has recently leveraged the model of "challenges" with prizes for the teams that can create the best open source solutions. For example, in collaboration with the Partnership on AI, Microsoft, and academics from University of Oxford, Cornell Tech, MIT, and UC Berkeley amongst others, Facebook designed and rolled out the Deepfake Detection Challenge (DFDC) aimed at the development of machine learning models that everyone can use to better detect when AI has been used to **alter a video in a misleading way**.[10] To support that research, we even went so far as to create a brand new data set of deepfakes based on videos we collected from paid actors who agreed to having their images manipulated for this work. Similarly, we launched the **Hateful Memes Challenge**,[11] an online competition to spur faster progress across the industry in dealing with hateful content and help advance multimodal machine learning more broadly. In support of that effort, we released the first data set designed specifically to enable research into multimodal hate speech.

Focusing on AI for Good.

AI can be a powerful force for good, driving economic growth and helping address climate and health issues amongst others. That's why we are collaborating with the Digital Ethics Lab of the University of Oxford to assess, map and explore how AI can help meet the **United Nations Sustainable Development Goals (UN SDGs)**.[12] Meanwhile, In response to the COVID-19 pandemic and as a part of Facebook's Data for Good programme, Facebook has acted quickly to offer interactive maps on population movement that researchers can use to understand the effectiveness of quarantine restrictions, using aggregated data to help **protect people's privacy**.[13] In the EU, Facebook Ireland also partnered with the University of Maryland to launch a voluntary symptoms survey designed to help health researchers identify COVID-19 hotspots earlier, and we are using aggregated data from that survey to generate more interactive maps that we plan to update daily throughout **the outbreak**.[14] And, of course, AI has played an essential role in helping us identify disinformation about the pandemic so that we can remove it from our platforms.

That is only a brief summary of all of the work that Facebook is doing to foster excellence in the field of AI, and we now stand ready to begin a new phase of that work: constructively collaborating with the European Commission on the future of AI governance.

B. BUILDING TRUST IN AI

Facebook supports the Commission's overall goal of building an "ecosystem of trust" in AI through forward-thinking approaches to governance, recognising the need to ensure that AI systems making important decisions are fair, transparent, accountable, and privacy-respecting. Finding the right balance in regulation is especially challenging when considering such a novel technology that is constantly evolving – but Facebook is committed to collaboratively arriving at the right answers.

A general consensus has developed in the past several years around fundamental Responsible AI principles rooted in **human rights and consumer protection**.[15] The European Commission's own **Ethical Guidelines for Trustworthy AI**,[16] which are the basis of the White Paper recommendations and have been helpful in guiding our own internal AI frameworks and practices, alongside the OECD AI Principles, to which the Commission and several EU Member States have greatly contributed, have helped define this emerging global consensus.

Facebook itself has been working hard to translate these principles into practice. To drive this work, we have created a dedicated, multidisciplinary **Responsible AI (RAI)** team of ethicists, social and political scientists, policy experts, AI researchers and engineers focused on understanding fairness and inclusion concerns associated with the deployment of AI in Facebook products. That team's overall goal is to develop guidelines, tools and processes to tackle issues of AI responsibility and help ensure these resources are widely available across the entire company so that we have a systematic approach to these hard questions.

Below are just some examples of the work that Facebook's RAI team and other product teams have been focusing on to further improve the fairness and transparency of our AI systems:

- Facebook built **Fairness Flow**,[17] a technical toolkit for detecting specific kinds of statistical bias in certain machine learning systems. In particular, Fairness Flow can be used to determine whether binary decisions made by algorithms (or by humans labelling training data) differ in accuracy for different subgroups of people. For example, Fairness Flow can test whether our hate speech classifier – the algorithm that predicts whether a piece of content is hate speech under our Community Standards – is as accurate for men as it is for women, or for younger people as it is for older people. Although testing of more complex machine learning systems will require the development of more complex tools, Fairness Flow has already helped us address important real-world risks. For example, during the 2019 India general elections, in order to assist human reviewers in identifying and removing political interference content, Facebook built a binary model to identify high-risk content that discussed civic or political issues. We used the Fairness Flow tool to help ensure that the predictions of that classifier model as to whether content was civil/political were equally accurate across languages and regions in India. This is important because if the model had systematically underestimated risk for content in a particular region or language, then fewer human review resources would have been allocated to that region or language than necessary.

- When we built the automatic camera for our Portal smart video calling device that can centre users in the camera's frame as they move around, we realised that this movement feature didn't work as well for certain genders and skin tones. In order to help correct this problem in Portal, we developed **Inclusive AI** guidelines that specify how to build representative test data sets across **different skin tones and genders**.[18]
- Facebook has begun to pilot an internal **Fairness Consultation Process**, managed by a core group of employees with expertise in fairness in machine learning, privacy, and civil rights, that helps product teams understand potential issues and connects them to additional subject matter experts in law, policy, ethics, and ML fairness. As part of the pilot, guidance was developed to help product teams and their cross-functional partners spot potential issues with AI fairness and flag them for additional input and discussion by the consultative group. The goal is to take learnings from this voluntary consultation process to inform the development of an overall fairness approach across the full range of Facebook products.
- In the realm of transparency and explainability, we have built a range of **"Why Am I Seeing This?" (WAIST) features** to explain to users why our automated systems are showing them particular content, including "Why am I seeing this post?" to help users better understand how automated content ranking works in **News Feed**,[19] and "Why am I seeing this ad" to help explain **ad targeting**.[20] We are also innovating around how to be even more transparent to users about the data that drives our systems. **Off-Facebook Activity** lets you see a summary of the apps and websites that send us information about your activity, and **clear this information from your account if you want to**.[21] And our tools **Download Your Information** on Facebook, and **Download Your Data** on Instagram, consistent with the right to data portability that GDPR introduced to enhance innovative competition, allow people to easily view and download data that they've shared through **our services and data about them that are stored by our services**.[22, 23]

Facebook and other AI developers are continually exploring new tools and processes for enhancing the fairness and transparency of our AI systems in line with emerging Responsible AI principles. However, there is still a general lack of consensus on exactly how best to translate these broad responsible AI principles into practice, especially across the wide range of different contexts in which AI can be deployed, and approaches to these problems are still evolving. Dialogue is still ongoing around hard questions like how best to guarantee algorithmic fairness, advance AI explainability, or test and document data and model quality.

- Facebook is doing its part to contribute to the dialogue around emerging best practices across industry and the broader policy community. This includes deep participation in important multi-stakeholder fora like the **Partnership on AI**,[24] of which we were a founding member, and support for the growing academic community focused on issues of AI governance and ethics, including the aforementioned **TUM Institute for Ethics in AI**. Facebook has also been deeply involved in ongoing international efforts to which the Commission and several Member States have also been decisively contributing, like the OECD's AI work. Facebook is a part of the expert group that helped formulate the OECD Principles on AI and is now working with the OECD Network of Experts on AI (ONE AI) within the **OECD's AI Policy Observatory** to help define what it means to implement these **principles in practice**.[25]
- Facebook is also building collaborations with regulators themselves to imagine the future of AI regulation. For example, Facebook recently launched a **Policy Prototyping programme** on AI explainability to bring technology regulators and industry actors together to co-create, refine, and test AI governance frameworks using regulatory sandboxes. Our first policy prototyping collaboration is with **Singapore's Infocomms Media Development Authority and Personal Data Protection Commission**, and we are exploring other ways to similarly engage with policymakers on emerging AI policy questions in the EU and around the world, including a series of **"Design Jams"** through Facebook's TTC Labs initiative focused on topics like **algorithmic transparency**.[26]

Facebook believes that any regulation of AI should support and build on these ongoing efforts to establish best practices in the field of responsible AI development, rather than risk cutting them short by prematurely codifying inflexible mandates in this fast-changing field. Legal certainty around AI developers' obligations can be achieved while still preserving the flexibility to accommodate changing needs and norms – and the ability to take full advantage of the powerful economic benefits of AI – as the technology evolves. It is in that spirit that we offer the following recommendations about how the Commission's AI proposal could be clarified.

03

Recommendations

A. CLARIFYING DEFINITIONS AND SCOPE

Facebook appreciates the Commission's recognition that any new regulation must be carefully tailored to address the most high-risk applications of AI or else risk significantly burdening AI-driven innovation and economic growth. A clearly-defined risk-based approach is required to avoid a blunt "one size fits all" regulation that unnecessarily reaches across the wide and diverse range of AI uses. A clearly defined risk-based framework will also help ensure that regulatory intervention is proportionate and not overreaching. A clear definition of regulatory scope in turn requires clear definitions of what counts as "high-risk" "artificial intelligence."

1. DEFINING AI

Facebook agrees with the Commission on the importance of arriving at a definition of “Artificial Intelligence” that is “sufficiently flexible to accommodate technical progress while being precise enough to provide the necessary legal certainty,” while we also recognise the challenges of doing so.

We are eager to participate in continuing dialogue on how best to do that and to help ensure that the Commission clarifies its intentions on this score, since the White Paper’s general reference to “technologies that combine data, algorithms and computing power” threatens to encompass nearly all modern software systems. However, just as a definition should not be too broad, it should also not be too narrowly focused on a detailed and prescriptive description of the underlying technical elements of AI and machine learning. These are dynamic and continuously evolving fields, and any attempt to encapsulate their technical details will inevitably and rapidly become outdated. Therefore, we urge the Commission to take two important themes into consideration as it evaluates potential definitions.

A Focus on Software that Learns.

Many software systems make decisions. What makes newer AI technologies unique, and what raises unique governance questions around fairness and accountability, is that modern AI systems *learn* to make decisions over time – up to and including complex sets of decisions around human-level cognitive acts like driving a car, playing chess, or making judgments about someone’s job application – rather than their decisions being based on hard-coded rules. Any definition needs to focus on this aspect of AI or risk sweeping in a broad range of technologies that do not raise these concerns and/or are already subject to extensive regulation.

A Focus on AI in Context.

The Commission is not seeking to regulate AI because it is AI, but because AI when used in certain sensitive contexts raises unique risks *for people*. The primary focus of any definition of AI should therefore focus less on how this technology operates and more on who is impacted by it and how. In this respect, it may be useful for the Commission to look to GDPR’s human-centered treatment of automated decision-making, which focuses on systems that make decisions about people or that otherwise impact them in meaningful ways – *i.e.*, that have legal or similarly significant effects – rather than on technical specifications. Such a context-aware, human-centered focus would be particularly appropriate here considering the Commission’s commitment to a risk-based approach to regulation.

Considering these themes, one definition of AI worth giving special consideration is the one offered by the Expert Group on Artificial Intelligence at the OECD:

> An AI system is a machine-based system that is capable of influencing the Environment by making recommendations, predictions or decisions for a given set of Objectives. It does so by utilising machine and/or human-based inputs/data to: i) perceive real and/or virtual environments; ii) abstract such perceptions into models manually or automatically; and iii) use **Model Interpretations to formulate options for outcomes**.[27]

This definition is notable in that it addresses the need to focus on learning systems, while also being mindful of the specific context of different parts of the AI development lifecycle. The individual elements of this definition, following AI through its various development phases, were devised to map to and help implement the OECD's AI Principles which share many commonalities with the Commission's own Trustworthy AI Ethical Guidelines. Were the Commission to offer a definition with a similarly granular view, that definition could support more detailed guidance around the specific risks and mitigations most appropriate to the different stages of AI development.

In other words, such a definition would allow for more nuanced AI governance focused on specific problems arising in specific contexts – which is the same goal that drives our following suggestions on how to initially define, and ultimately assess, the level of risk posed by specific AI systems.

2. DEFINING HIGH-RISK AI

Facebook is aligned with the Commission's goal of limiting regulation to those highest risk AI uses that require it. We appreciate the Commission's attempt to do so by proposing only to regulate AI uses that meet two cumulative criteria: falling within a specific list of sectors where risk is most likely and being used in a manner that "significant risks" are likely to arise. However, we urge greater clarity in defining those criteria to ensure both legal certainty and proportionate scope, and we are eager to participate in future dialogue around how best to shape those criteria. For now, we would like to highlight three key recommendations:

Ensure Precision Around Sectors.
In order to avoid disproportionate burden on businesses both large and small, we urge that specific market sectors and sub-sectors be clearly and narrowly defined. If the goal is to accurately reflect those areas of the market where high risk is most likely, categories as broad as "transport" or "health" are of limited use. Such broad categories risk unnecessarily capturing broad swaths of the economy and requiring countless businesses to then expend legal and technical resources evaluating whether their AI uses meet the second criterion.

Remove "Immaterial Damage" from the Definition of High-Risk Uses.
Building on and seeking consistency with GDPR Art. 22's approach to high risk in the context of automated decision making as the White Paper does – that is, focusing on the legal and similarly significant effects of an AI system on people – has a number of benefits. Most importantly, as it is a pre-existing definition supported by guidance already provided by Art. 29 Working Party and endorsed by the **EDPB**,[28] it provides greater legal certainty than a new definition and would also be easier for organisations to operationalise as they already refer to it as part of their GDPR compliance. It is also reasonable to propose that AI uses posing a *direct* risk of injury, death, or *serious* material damage also be included. However, "immaterial damage" is so broad and vague that it could easily encompass an unreasonably wide set of circumstances. If there are particular kinds of immaterial damage the Commission believes would not already be covered by the inclusion of AI uses with legal and similarly significant effects, then those should be specified, rather than creating a situation where neither companies nor future regulators will have clarity about whether their AI systems are covered by the regulation.

Remove the "Exceptional Instances" Clause.
Codifying vague exceptions based on unspecified criteria risks completely eliminating the legal certainty that regulation is intended to provide and threatens to enable regulatory overreach. Any concern that the regulation will not sufficiently cover the necessary range of high-risk uses should be

addressed by defining the relevant sectors and the concept of “high risk” with sufficient care, and by issuing clear guidance with specific examples of high-risk uses (as described below).

Again, Facebook supports the goal of the Commission’s proposed definitions and is eager to continue the dialogue around how they may be refined. Ultimately, however, these definitions must be supplemented by in-depth assessment of specific uses in specific contexts and considering the AI benefits against which the AI risks are being balanced. One potential method for doing so is outlined below.

B. ALIGNING WITH GDPR'S APPROACH TO RISK ASSESSMENT AND ENFORCEMENT

1. The Costs of *Ex Ante* Conformity Assessments

Facebook supports the Commission's overall goal of building a thoughtful, proportionate, and risk-based approach to AI governance. However, the question of AI regulation is not a binary one, but rather a question of scope and calibration between hard law, co-regulation, and soft law instruments. Particularly in an area of technology that is changing as quickly as AI, we need a regulatory structure that can dynamically recalibrate based on evolving AI best practices, otherwise we risk unnecessarily stifling beneficial innovations.

The structure proposed in the White Paper lacks that flexibility. The imposition of prior conformity assessments based on the mechanical application of a "high-risk" definition, uninformed by the specific context of the AI use at issue, would be a costly, time-consuming, and inflexible approach. It would also be quite disproportionate, at least outside of the context of the very highest-risk systems like those that risk direct physical injury or death.

Indeed, such an inflexible regime could unnecessarily delay implementation of AI technologies addressing urgent human needs. For example, Facebook's recent work to quickly update its content integrity systems to address misinformation around COVID-19 would have been impossible if substantial changes required prior approval from a regulator or a certifying third party. The same thing can be said for other AI applications co-developed by Facebook and external stakeholders in the context of the fight against COVID-19, such as the use of AI to predict the spread of the virus so emergency responders and hospitals can better allocate their resources, or the use of AI to translate COVID-related educational materials into languages that would otherwise be underserved.

More generally, and considering AI's increasingly central role to human productivity, the economic and innovation costs of such a broad prior conformity requirement could also hinder the EU's economic growth and its ability to recover from economic shocks like the one we are all facing now. In order to assess these risks in line with the Commission's Better Regulation Guidelines for best practice, and as it did with GDPR, we urge the Commission to undertake a full impact assessment including an economic impact assessment in regard to its proposal.

We urge the Commission to look to the precedent of GDPR in another way, as well.

2. An Alternative: Self-Assessment of Risk as in GDPR

The Commission recognised the costs and inflexibility of a prior conformity assessment approach when, in GDPR, it consciously abandoned the *ex ante* external "prior checking" system existing under Directive 1995/46/EC. Through GDPR, the Commission instead established the duty to implement accountable data protection programmes that include Data Protection Impact Assessments (DPIAs): *ex ante* self-assessments for data processing likely to be high-risk.

A similar approach to AI – requiring "Automated Decision-making Impact Assessments" or ADIAs, akin to DPIAs – would be a more balanced alternative to requiring blanket prior reviews by a regulator of all "high-risk" AI applications as the White Paper recommends. Such an approach would also align with the GDPR's principle of accountability whereby organisations acting as controllers are in the best position to assess, determine, and document the level of risk raised by their own processing activities, and to mitigate those risks accordingly.

Although much would turn on the details of how such an ADIA process was implemented, it is not hard to imagine an approach that follows in GDPR's footsteps and achieves the Commission's desired goals in a much more flexible and less costly manner than prior conformity assessments. For example, the following features would be consistent with GDPR's approach while also addressing some of its shortcomings:

- **Codifying an ADIA Process**. The Commission could require that AI systems likely to meet its criteria for high-risk be subjected to an ADIA – a step-by-step self-assessment process for identifying and quantifying risks and benefits and identifying and documenting mitigations for those risks. Under such a structure, measures described in the White Paper as mandatory could be integrated directly into the risk assessment process. For example, the implementation of appropriate human oversight or the taking of concrete steps to assure appropriate quality and breadth of training data would work as mitigations that reduce the overall risk. The basic components of the ADIA process could be formalised at a high level in the regulation, leaving room for evolving soft law instruments to provide more detailed guidance.
- **Providing Guidance on the ADIA Process**. The Commission, ideally collaborating with a standing committee involving diverse expert stakeholders from the private and public sectors, could provide its own detailed guidance that it can update when needed. Such guidance could be a detailed taxonomy of the kinds of risks and harms caused by

automated decision-making systems, indicative examples of AI uses that are presumed to be high-risk (a presumption that could be rebutted with appropriate mitigations documented in an ADIA), and a methodology that developers could follow when seeking to identify and quantify harms. When considering such a methodology, we suggest consulting **Singapore's IMDA Model AI Governance Framework**,[29] namely the matrix described in that document to help organisations determine the level of human involvement in AI decision-making. That matrix lays out a number of factors that could be used as operational guidance to assess risk in the context of ADIAs, such as (a) probability of harm; (b) severity of harm; (c) the nature of the harm; (d) the reversibility of harm, which depending on the context, could include the ability for individuals to obtain recourse; and (e) whether it is operationally feasible or meaningful for a human to be involved in the decision-making process.

- **Considering Context and Aligning Approaches**. Such detailed guidance from the Commission would ensure adequate consideration of the specific context of the automated decision-making at issue – context that is lacking in any blunt definition of “high risk” – and would ensure a focus on concrete and measurable harms. This taxonomical and methodological approach would also help avoid repeating the difficulties that have been experienced around achieving alignment between Member States on a consistent approach to **risk assessment under GDPR**.[30]

- **Weighting AI Benefits, Not Just Harms.** Ideally, this AI risk assessment process could focus not only on harms but also on the benefits of the AI system at issue. By focusing also on the positive effects of AI, one would also take into account the risks of *not* using AI in a particular context, along with the beneficial uses and applications of AI, such as their potential to reduce or mitigate discrimination, increase human autonomy, enable scientific discoveries, or contribute to higher standards of living. This would enable a more holistic and comprehensive risk assessment procedure, based on a calculation of both benefits and costs.

- **Leveraging Soft Law to Maintain Adaptability.** Consistent with GDPR's approach to complementing DPIAs with approved codes of conduct as a way to assess the impact of the processing operations performed by controllers or processors (art. 40 and art. 35.8), ADIAs should be complemented and further detailed in industry best practices, codes of conduct, codes of practice, standards and industry-led certification mechanisms. By deferring the details of the specific assessment and documentation requirements to such soft law instruments, as GDPR does, the Commission could ensure that the regulation is able to adapt as the technology and expectations around it continue to evolve. In doing so, it could support and draw from ongoing dialogues around what it means to develop and deploy AI responsibly, such as those occurring through the OECD's AI Policy Observatory supported by the multi-stakeholder **OECD network of experts on AI (ONE AI) of which Facebook is a part**.[31]

- **Reserving Prior Consultation with the Regulator and *Ex Post* Auditing for Exceptional Cases.** Consistent with GDPR's approach, an AI regulation could require prior consultation with the relevant regulator only when the ADIA process has resulted in the identification of residual high risks for which appropriate mitigations are not reasonably available or have not been identified. This would encourage organisations to proactively consider and adopt mitigations that reduce the initial high risk to an acceptably low level. Meanwhile, any external audit requirement could be shifted to the *ex post* enforcement phase, reserved for situations where the developer did not conduct an ADIA when required, the ADIA was manifestly incomplete, or to verify that any additional safeguards indicated by the competent authority in a prior consultation were implemented. This change would enable organisations to offer their products and services in the EU without routine mandatory third-party inspections and assessments that would create unnecessary hindrances to speedy technological innovation. This change would also alleviate the oversight and enforcement burden that supervisory authorities would otherwise need to shoulder, while still keeping organisations accountable.

We offer the above framework for a potential ADIA-centered regulatory regime as an example to demonstrate that there are multiple ways for the Commission to build on existing regulatory requirements, and strike a more sustainable balance between the need to govern and the need to preserve AI innovation and economic growth in the EU, while also supporting the continued evolution of best practices in this area.

In contrast, a more prescriptive regulatory approach, articulating in detail a set of mandatory requirements for the design, development and deployment of AI systems in "hard" law, would undermine the creation of these agile soft law instruments and could constrain the development of AI within the EU. This would result in a static and inflexible regulation, populated by a set of requirements that would be outdated quickly, failing to accompany the pace of technological evolution and, as such, failing to provide useful and meaningful guidance to AI developers. We wholly agree with the Commission's statements in the White Paper that "[a]s a matter of principle, the new regulatory framework for AI should be effective to achieve its objectives while not being excessively prescriptive so that it could create a disproportionate burden, especially for SMEs," and that "[g]iven how fast AI is evolving, the regulatory framework must leave room to cater for further developments."

Questions and Tensions to be Resolved 04

Having addressed the overall scope and structure of the White Paper's proposal, we turn to a more fine-grained consideration of several of its proposed mandatory requirements and certain aspects of its proposed scope and enforcement. These elements of the proposal raise a great many novel risks, questions and tensions that will need to be addressed as the legislative dialogue proceeds. Below, we elaborate on the most concerning issues that we perceive based on our own experience dealing with the day-to-day operational and legal challenges of a large-scale AI developer, and in some cases, we outline possible approaches to resolving them. We look forward to continued constructive engagement with the Commission on these hard questions.

A. CONCERNS REGARDING OVERALL SCOPE OF PROPOSED NEW LEGAL DUTIES

1. Risk of Overlap or Conflict with GDPR

The White Paper outlines a long and prescriptive list of proposed mandatory requirements on record-keeping, ML process documentation, data set attributes and composition, robustness and accuracy of models, human oversight obligations and specific biometric identification safeguards. However, several of these requirements or elements of them may duplicate obligations that are already present in GDPR. Therefore, there is a need for a clear distinction between the requirements that AI developers already need to comply with under the GDPR and those that go beyond GDPR.

A clear assessment should be made of what mandatory requirements the Commission is considering that are already included in GDPR, and the Commission should carefully avoid creating another potentially conflicting AI data regime or one that would impose compound liability for the same conduct. A similar inquiry to avoid double-regulation must also be made in regard to AI uses in already heavily-regulated sectors like health and transport.

2. Lack of Clarity Around Retrospective vs. Prospective Enforcement

The White Paper does not answer a critical question for every AI developer: Would regulatory requirements apply to models already in production where the information for the required documentation may not exist? Would developers be forced to delete and start over? How long would they have to do so? To avoid unnecessarily burdening developers and creating a massive bottleneck in terms of testing centres (to the extent third party testing is required), we would urge the Commission to clarify that AI systems in production prior to enforcement of the regulation would only need to be assessed for conformity if and when they are substantially updated or modified.

3. Risks Arising from Proposed Modifications to Existing Product Liability and Safety Laws

In addition to the AI regulatory proposal contained in the White Paper, the Commission has suggested modifications to existing safety and product liability laws, including the General Product Safety Directive and the Product Liability Directive (both of which are technology neutral), to account for potential harms caused by AI-powered systems. Facebook has some concerns about certain of these proposed modifications and welcomes more in-depth discussions with the Commission on this in the future. For example, at a high level, it is unclear whether such an expansion of existing regulation is needed to address the potential challenges arising from AI, IoT or robotics identified in the White Paper. Without more in-depth consideration of whether existing regulation is adequate, there is an increased risk that any proposed modifications will not provide clarity, but instead create areas of legal uncertainty, including an increased risk that AI developers could face conflicting and concurrent liability under a patchwork of various regulatory regimes. Relatedly, depending on how any expansion of these laws is implemented, AI developers could be subject to varying and inconsistent definitions of “AI,” “harm,” or “risk” arising from different Member State enactments, resulting in additional uncertainty and confusion.

As another example of where more clarity is needed, the White Paper raises open questions around the assignment of responsibility where there are multiple parties involved in the development and deployment of an AI system. Specifically, the white paper proposes “requesting cooperation between the economic operators in the supply chain” with little clarity on how such an approach would, in practice, resolve the question of responsibility. Such looming uncertainty could result in deterring innovation, such that AI developers may discontinue, or never begin, development of systems in this space given the potential liability they may face.

With respect to any proposed expansion of existing liability frameworks to AI systems, Facebook questions the application of strict liability principles, which are more often reserved for abnormally dangerous activities or product defects resulting in physical injuries, without further consideration of whether such systems warrant such an inflexible approach. Such an expansion of existing liability laws would also likely significantly impact innovation in AI, as well as investment in AI systems, and should be given careful consideration.

B. TENSIONS BETWEEN THE PROPOSAL AND OTHER IMPORTANT VALUES AND OBLIGATIONS

1. Tensions with the Goal of Data Protection

We are concerned that the proposed requirement to store training data sets in certain cases could create a direct tension with data minimisation and data retention policies intended to protect users' data, which generally err on the side of reducing data retention and giving people control over their data when it is stored. Similarly, this requirement could unintentionally prohibit privacy-preserving AI approaches like federated learning, where the data used to train an algorithm is not held in a centralised location but is instead distributed across multiple devices such as a user's smartphone. By keeping all training data on other devices rather than uploading it to a central provider, these systems offer much greater protection for users' data. A requirement to retain and potentially share a consolidated and centralised version of the training data would undermine the privacy promise of federated learning.

Another significant data protection tension arises when considering the White Paper's treatment of issues of fairness and bias. The Commission notes that ensuring fair and unbiased or non-discriminatory systems may require the use of additional information to ensure that data sets 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." This requirement highlights a tension between the value of fairness, translated into the need to use sensitive demographic data in order to test data sets and models for potential bias, and the value of data protection, which requires additional safeguards and conditions be applied to the collection and processing of sensitive data.

This tension is highlighted by the general prohibition under the GDPR against processing sensitive categories of data. The exceptions to that prohibition have traditionally been construed very narrowly, although not consistently – different Member States have their own readings of these exceptions and are also allowed to impose additional local restrictions on sensitive data processing. Therefore, we would welcome clarity from the Commission on how to resolve this legal tension between data protection and algorithmic fairness at the EU level.

As it weighs how to resolve this tension, we recommend that the Commission consider these questions:

- What legal structures enable sensitive data to be obtained and used for the specific purpose of measuring algorithmic bias, and in what circumstances? Could this use derive from the legal duty under EU law (art. 21 of the EU Charter of Fundamental Rights of the European Union) to avoid discrimination?
- What categories of sensitive data should private organisations obtain and use? And who should decide on these categories? What controls should be implemented to ensure that these categories of sensitive data are only used for this specific purpose unless otherwise authorised?
- To what extent would pseudonymisation of sensitive categories of data be considered a sufficient safeguard to mitigate data protection concerns and help foster algorithmic fairness?
- Should the regulatory proposal differentiate between having access to sensitive data to measure and test models for bias, and including that data in the models themselves in order to ensure unbiased outcomes?
- When, if ever, would it be appropriate to infer sensitive data about people for the purpose of testing for discrimination? What if it was otherwise impossible to obtain it?
- How should people be informed about the use of this type of data for bias measurement or mitigation? Should people be able to opt out to the collection and use of sensitive data even if it has been determined to be for a legitimate purpose? What if this opt-out makes it difficult or impossible to assess or correct potential bias?

These questions have broad implications for any company or organisation developing and deploying AI systems, and require broad conversations with stakeholders and policymakers in order to align on an approach to measuring and addressing prohibited discrimination. Notably, the Partnership on AI has convened industry-wide conversations and launched a **dedicated research project on this important topic**,[32] and Facebook is participating in that conversation as well as consulting with a diverse group of stakeholders on how to make progress in this area. The Commission's legislative process is another valuable venue for dialogue around this hard question.

2. Tensions with Data Security and Intellectual Property Rights

The technical and process modifications necessary to enable regular access to data and algorithms by external stakeholders would inevitably increase the likelihood of data breaches and of bad actors gaining unauthorised access to these data sets and the algorithms, which could then facilitate consequent abuses or misuses. One example of potential abuse would be adversarial attacks perpetrated by bad actors to gain access to training data sets used for terrorism-related content detection. By gaining access to those training data sets and learning how the model behaves, they could learn how to modify their content so that it would avoid being detected by our algorithms. Extensive information disclosure requirements around the details of model operation could risk creating similar security vulnerabilities.

Requiring AI developers to disclose their algorithms and the underlying training data to external parties for review could also compromise intellectual property protections and trade secrets for those assets. Companies invest significant resources over many years in developing the algorithms and identifying training data that creates more useful AI systems. Regulations requiring disclosure of this intellectual property could reduce future investments in this area, thereby dampening or stifling innovation in this field.

C. PRACTICAL QUESTIONS AND CONCERNS AROUND PROPOSED MANDATORY REQUIREMENTS

1. Practical Concerns Around Proposed Training Data Storage Requirement

The proposed requirement to retain training data sets "in certain justified cases" seems to assume that these data sets, and the models trained on them, are static. This is hardly ever the case. Given the growing number of applications of AI, size of data sets and sophistication in the training procedures, it may be technically impossible for organisations to keep all versions and iterations of training data sets ever used in designing and developing their AI systems.

We think that this proposed requirement regarding retaining data sets should be removed and would instead focus on developing the Commission's more general requirement that developers of high-risk systems document the procedure through which they created the data sets, along with information explaining what these data sets represent. This would require exploring a standardised way to keep metadata about training data sets, but not keep the data sets themselves. Examples of evolving approaches in this area include the Partnership on AI project "ABOUT ML" (**Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles**),[33] which is bringing together a diverse range of perspectives to develop, test, and implement machine learning system documentation practices at scale, and the "**Data Nutrition Project**",[34] which assesses data sets based on standard quality measures that are both qualitative and quantitative.

Importantly, the previously mentioned data protection and security tensions around storing and disclosing training data sets could be alleviated by focusing on test data and benchmarks, which can be a much more powerful tool for assessing the properties and behavior of a model. Verifying the properties of a training set is not a sufficient condition for ensuring a trained model will behave as desired when it makes automated decisions or predictions. Test sets are the ultimate tool for verifying the behavior of a trained model that will be effectively deployed. While the developer could create and retain its own test sets, external organisations could also create benchmark test sets (these could be synthetic but not necessarily) with the right statistical properties for the purpose of understanding whether a trained AI will behave as expected. This would foster even greater external accountability and a growing set of shared resources for cross-industry testing. One example of this sort of external data set creation is the work done by the Gender Shades project at MIT to create face databases with a focus on **greater inclusivity and representation in terms of skin colour and gender**.[35]

2. Ambiguity Around Proposed Robustness and Accuracy Requirements

The extensive list of proposed mandatory requirements for high-risk systems include ensuring that AI systems are robust, accurate, and that their outcomes are reproducible. These requirements need to be further clarified. It is necessary to define what is meant by reproducible, clarifying whether that means that a different model trained on the same data would have the same results, or that the same model deployed on new data would have comparable results. It is also important to define whether these information disclosure requirements are focused on reproducibility of individual predictions or decisions, or at an aggregate statistical level. While the latter is feasible, the former is extremely difficult to operationalise because AI systems are stochastic in nature. For example, standard training procedures like stochastic gradient descent (SDG) rely by design on a noisy process. Multiple training runs on a fixed training data set will result in models that have differences between them, even if their predictions are similar in a statistical sense.

In addition, the proposed requirement that AI systems must be able to adequately deal with errors or inconsistencies during all life cycle phases may be impracticable, as it is extremely difficult to address every possible error or inconsistency in these types of systems. Given how broad and technically challenging these requirements are, we recommend framing them as "reasonable efforts" requirements – for example, it is reasonable to expect average error rates to be within certain bounds. The Commission appears to endorse such a view when it notes, in the context of technical accuracy, only that "[a]ll *reasonable* measures should be taken to minimise the risk of harm being caused," but this should be made explicit in any final language.

3. Questions and Concerns Around the Proposed Requirement to Retrain AI Systems in the EU

Under the prior conformity assessment framework proposed by the European Commission, AI developers could be required to retrain their AI systems in the EU when requirements related to data training are not met. The purpose and method of this proposed enforcement requirement is unclear. Would these AI systems need to be retrained within the borders of the EU, and/or with EU data sets, and if so, with what data and in service of what goal? Speech recognition algorithms trained to recognise many different voices and accents, and computer vision algorithms trained to recognise objects or categorise scenes in images and videos, would be of much lesser quality and usefulness if their training data sets were restricted to "European data". If the goal is to address unintended bias, an overreliance on EU data could itself introduce its biases. Facebook is committed to exploring innovative approaches to detecting and mitigating statistical biases in AI, but a simple mandate to "use EU data" is not a solution.

Furthermore, improving the accuracy or behavior of a model isn't solely dictated by the training data, and imposing restrictions on the training would likely not be sufficient. Instead, as mentioned above, the focus should be placed on the test data, and on generating representative benchmark data sets, including data sets with adequate representation from the EU population.

Conclusion 05

As shown by the tensions and concerns highlighted above, AI poses novel and complex challenges to existing legal frameworks, and deciding what an effective, proportionate, and technically feasible AI regulation should look like will not be easy. There are many difficult questions ahead of us, but we are eager to collaborate on how to answer them, both in the context of this legislative dialogue and in other fora.

As noted above, Facebook is especially interested in the potential for using regulatory sandboxes and policy prototyping programmes to develop and test policy ideas around emerging technology like AI in collaboration with local regulators, as we are currently doing with Singapore. We would welcome such collaborations with the Commission or member state DPAs to develop, experiment and assess the impact of different evidence-based policy and regulatory approaches to AI before they are enacted in legislation. In the meantime, we look forward to engaging in conversation with policymakers and other stakeholders to exchange views on the Commission proposal and the challenging AI policy questions it poses.

Facebook values the Commission's leadership in this area and appreciates the opportunity for continued dialogue on these important issues. Together, we hope to ultimately chart a course for the future of AI governance that leads to a safer, fairer, and more prosperous society, with Europe as a global leader in AI innovation.

End Notes 06

1. See Mark Zuckerberg, *Big Tech Needs More Regulation*, Financial Times (February 16, 2020), **https://www.ft.com/content/602ec7ec-4f18-11ea-95a0-43d18ec715f5**

2. See European Commission, *White Paper on Artificial Intelligence - A European approach to excellence and trust*, COM (2020) 65 final (February 19, 2020), **https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf**

3. See *Facebook Community Standards*, **https://www.facebook.com/communitystandards/** (last visited June 11, 2020)

4. *Facebook Artificial Intelligence Research*, **https://ai.facebook.com/**

5. See *A look at Facebook AI Research Paris* (March 2017), **https://research.fb.com/blog/2017/03/a-look-at-facebook-ai-research-fair-paris/**

6. *PyTorch*, Facebook AI, **https://pytorch.org/** (last visited June 11, 2020)

7. *Institute for Ethics in Artificial Intelligence* (IEAI), **https://ieai.mcts.tum.de/** (last visited June 11, 2020)

8. See *Ethics in AI Research Awards - India* (June 17, 2019), **https://research.fb.com/programs/research-awards/proposals/ethics-in-ai-research-awards-india/**; and *Ethics in AI Research Initiative for the Asia Pacific request for proposals* (December 8, 2019), **https://research.fb.com/programs/research-awards/proposals/ethics-in-ai-research-initiative-for-the-asia-pacific-request-for-proposals**/

9. See *CeTys, GuIA.ai Inteligencia Artificial en América Latina y el caribe: Ética, gobernanza y políticas*, **https://guia.ai/** (last visited June 11, 2020)

10. See *Creating a data set and a challenge for deepfakes*, Facebook AI (September 5, 2019), **https://ai.facebook.com/blog/deepfake-detection-challenge/**

11. *Hateful Memes Challenge and Data Set*, Facebook AI, **https://ai.facebook.com/hatefulmemes** (last visited June 11, 2020)

12. *Oxford initiative on AIxSDGs*, **https://www.aiforsdgs.org/** (last visited June 11, 2020)

13. *Facebook Data for Good*, Facebook, **https://dataforgood.fb.com/** (last visited June 11, 2020)

14. Mark Zuckerberg, *How Data Can Aid the Fight Against COVID-19*, Facebook Newsroom (April 20, 2020), **https://about.fb.com/news/2020/04/symptom-surveys/**

15. See, *e.g.*, this chart mapping the commonalities between dozens of principles statements from governments, civil society, businesses and academia. **http://wilkins.law.harvard.edu/misc/PrincipledAI_FinalGraphic.jpg** (last visited June 11, 2020)

16. See High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI* (April 8, 2019) **https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai**

17. See Jerome Pesenti, *AI at F8 2018: Open Frameworks and responsible development* (May 2, 2018), **https://engineering.fb.com/ml-applications/ai-at-f8-2018-open-frameworks-and-responsible-development/**; and *F8 2018 Day 2 Keynote video* (May 2, 2018), **https://developers.facebook.com/videos/f8-2018/f8-2018-day-2-keynote/**

18. See *Building inclusive AI at Facebook* (May 1, 2019), **https://tech.fb.com/building-inclusive-ai-at-facebook/**

19. See Ramya Sethuraman, *Why Am I Seeing This? We Have and Answer for You* (March 31, 2019), **https://about.fb.com/news/2019/03/why-am-i-seeing-this/**

20. See *How does Facebook decide which ads to show me?* **https://www.facebook.com/help/562973647153813?helpref=faq_content** (last visited June 11, 2020)

21. See *What is off-Facebook activity?* **https://www.facebook.com/help/2207256696182627** (last visited June 11, 2020)

22. See *How do I download a copy of my information on Facebook?* **https://www.facebook.com/help/212802592074644** (last visited June 11, 2020)

23. See *How do I access or review my data on Instagram* **https://help.instagram.com/181231772500920** (last visited June 11, 2020)

24. *The Partnership on AI* (PAI), **https://www.partnershiponai.org/** (last visited June 11, 2020)

25. *The OECD Artificial Intelligence Policy Observatory*, **https://oecd.ai/** (last visited June 11, 2020)

26. See *Algorithmic Transparency Amsterdam*, **https://www.ttclabs.net/event/amsterdam** (last visited June 11, 2020)

27. See *Scoping the OECD AI principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)*, OECD Digital Economy Papers (November 2019 No. 291), **https://read.oecd-ilibrary.org/science-and-technology/scoping-the-oecd-27ai-principles_d62f618a-en#page7**

28. See Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, (2017), **https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053**

29. See Info-communications Media Development Authority (IMDA) and Personal Data Protection Commission Singapore (PDPC), *Model Artificial Intelligence Governance Framework*, 2nd Edition (2020), **https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf**

30. As of today, each EEA Member State has filed its own list of high-risk processing activities for the EDPB and 4 of them have also filed their lists of no-high-risk processing. The different types of processing varied quite substantially amongst these lists, which compelled the EDPB to issue 35 opinions, one to each DPA, in order to try to build common criteria. The opinions can be found at **https://edpb.europa.eu/our-work-tools/consistency-findings/opinions_en**

31. OECD Network of Experts on AI (ONE AI), **https://oecd.ai/network-of-experts** (last visited June 11, 2020)

32. See M. Andrus et al, *Working to Address Algorithmic Bias? Don't Overlook the Role of Demographic Data* (April 24, 2020), **https://www.partnershiponai.org/demographic-data/**

33. See *ABOUT ML - Annotation and Benchmarking on Understanding and Transparency of Machine learning Lifecycles*, **https://www.partnershiponai.org/about-ml/** (last visited June 11, 2020)

34. See *The Data Nutrition Project*, **https://datanutrition.org/** (last visited June 11, 2020)

35. See MIT Media Lab, *The Gender Shades project*, **https://www.media.mit.edu/projects/gender-shades/overview/** (last visited June 11, 2020)

FACEBOOK