OpenAI

17 April 2020

OpenAI
3180 18th St
San Francisco, CA 94110

Comments on European Commission White Paper:
On Artificial Intelligence -
A European Approach to Excellence and Trust

This submission from OpenAI provides feedback on the White Paper: *On Artificial Intelligence - A European Approach to Excellence and Trust*. We previously submitted feedback on the Ethics Guidelines for Trustworthy AI's Technical Robustness and Safety section; we will draw from our previous feedback throughout this response. We offer recommendations for the Commission to consider when taking action towards establishing guidelines and requirements for developing responsible AI. We hope this submission will help the EU develop its own technical capacity, and aid it in developing an EU-wide framework for AI regulation and Trustworthy AI. This feedback will first respond to the Ecosystem of Trust, then to Ecosystem of Excellence. If a section is not highlighted, we do not have specific, actionable feedback.

About OpenAI

OpenAI is an artificial intelligence research company based in San Francisco whose mission is to develop transformative artificial intelligence and ensure it benefits all of humanity. OpenAI's work is primarily built around three areas: technical capabilities research and development; AI safety research and development; and policy work, which supports informed AI policymaking. For the last year, we have been conducting multi-stakeholder research to support a paper titled "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims,"[1] which outlines actions that governments, organizations, and developers can take to improve the verifiability of claims made about AI systems. We will draw on insights from this report throughout this response.

[1] View report on arXiv (https://arxiv.org/abs/2004.07213) or at http://www.towardtrustworthyai.com/

Investing in Trust and Excellence

AI systems are embedded within society and have associated societal effects. They must therefore be evaluated as socio-technical systems. To properly consider the impact of such systems, governments and multinational governance groups, such as the European Commission, need to **invest in their own capacity to understand, analyze, and evaluate the technical traits of these systems as well as their impact in deployed environments**. Governments must invest in this essential capacity so they are easily able to create the oversight tools and regulatory interventions that best address citizens' needs while supporting further AI research, development, and deployment.

Much of the White Paper's stated goals require that the Commission and associated regulatory bodies be able to accurately assess AI technology. We recommend the Commission consistently seek opportunities to develop government and intergovernmental capacity to analyze the technical specifics of contemporary AI technologies, how the field is maturing, and where future progress is likely to occur. Such investments will enable tech-informed policymaking, and will help the Commission craft regulations that are well calibrated to technical AI progress. Such efforts will also make it easier for the Commission to support the development of common, practical standards for AI technology, while easing coordination among multiple European countries towards the outlined requirements.

An Ecosystem of Trust: Regulatory Framework for AI

In the AI ecosystem, there is no consensus around what makes an AI system trustworthy, nor is there consensus around how to fully achieve each of the requirements. We recommend the Commission develop common notions of what makes a system trustworthy. This will require standardized definitions of concepts like technical robustness and safety, which will require coordinating with academia and the private sector. The Commission's investing in its own technical capacities will give it a strong foundation from which to assess and define AI and AI-relevant concepts like trustworthiness; these definitions will help develop effective legislative approaches and make legislation flexible to AI advances and changes in the AI ecosystem.

A. Problem Definition

We suggest that regulation not focus solely on AI, but also include effective oversight of those (e.g., companies) who are deploying a given technology (e.g., AI). AI systems can cause material harm through improper functioning but they can also cause immaterial harm (e.g., loss of privacy, limitations to the right of freedom of expression). These immaterial harms are not innate to the technology of AI, but rather are a consequence of those who deploy AI. Those who deploy include companies or public sector organizations and fall under existing regulatory tools.

B. Possible Adjustments to Existing EU Legislative Framework Relating to AI

Limitations of scope of existing EU legislation

We recommend the Commission adopt a sector-specific approach to applying safety legislation to AI. For example, many consumers use free applications with AI systems that automatically modify their appearance for the purposes of entertainment (e.g., camera filters that superimpose rabbit ears on a person). These applications are low-risk compared to other use-cases, and should be treated with an equivalently light touch. By comparison, some sectors like finance and healthcare involve systems that have potentially much higher impacts and require sector-specific legislation.

Changing functionality of AI systems

We agree with the ideas in this section and encourage the Commission to consider a "spot check" approach where an agency regularly tests deployed AI systems to ensure they are functioning as intended. For example, developers may need to show that an image classification system satisfies a certain pre-agreed upon standard for avoiding egregious biases prior to deployment. Regulators should periodically check on applications once a system is deployed to see if it continues to meet regulatory standards. Today, many of these spot checks are conducted on an ad-hoc basis by interested individuals (e.g., in April 2020, independent researchers used spot checks to identify a racist output in a vision-classifier system developed by Google)[2].

Uncertainty as regards the allocation of responsibilities between different economic operators in the supply chain

We suggest liability be focused on the developer of the product; this incentivizes developers to create their own processes for ensuring the safety of AI systems deployed on their products.

Changes to the concept of safety

We agree with this point and encourage the Commission to increase resources allocated to the EU Cybersecurity Agency (ENISA) so that it can scalably analyze the changing AI and threat intelligence landscapes.

[2] See: Google apologizes after its Vision AI produced racist results (Algorithm Watch, accessed April 7th 2020).

C. Scope of a Future EU Regulatory Framework

As defined in the beginning of this White Paper, AI has three main elements: data, algorithms, and computing power. AI algorithms require both computing power and data to run. This section clarifies only two of the three main elements (data and algorithms), but all three should be addressed.

Defining Risk

AI is an omni-use technology where risks are not always immediately apparent, especially for systems with many use cases; the same system can be used in a low-risk sector, then quickly adapted to high-risk ones. For example, facial recognition can be used for creative mediums, but can also be repurposed for privacy-infringing acts of surveillance. We therefore encourage regulators to define risks at the levels of specific *sectors* rather than relating to specific technologies, and to use a sector-specific lens to define risk differently in different regimes.

D. Types of Requirements

Training data

We recommend the EU **fund the creation of representative datasets to address fairness in AI development.** The Commission should lead by example in funding and creating training datasets that can help address EU safety and non-discrimination requirements.

All actors in the AI ecosystem, from developers to users, can benefit from Commission-crafted datasets; few actors outside policy bodies have the incentive and resources to make representative training data available. Funding the creation of these datasets and making them accessible will not only give guidance on what training data should look like, but will also provide a form of soft governance by incentivizing researchers and developers to create more representative AI applications via the use of these datasets.

However, we should be careful to avoid treating representatives as a silver bullet for solving some of the issues inherent to AI development; datasets are difficult to tie to a specific desired AI outcome a priori. AI capabilities are typically developed through repeated cycles of dataset collection, training, and evaluation. It is therefore unlikely that developers or regulators can prescribe the types of data needed to train on to “cover all the relevant scenarios needed to avoid dangerous situations”; instead, they need to develop systems, identify their limitations, and use these limitations to guide further dataset creation.

Another challenge is determining what deems a dataset “sufficiently representative”. The Commission should fund dataset analysis research, focusing on techniques to better predict how a dataset might relate to a particular desired outcome (e.g., satisfying a certain threshold of operation for a representative demographic group). One specific way the Commission can

support the creation of more representative datasets is through the use of “bias bounty” programs. Bias bounties are similar to the “bug bounty” programs that are already popular in the information security industry; “bias bounty” programs provide a legal mechanism for reporting biases directly to AI developers. Bounty programs could lead to increased emphasis on documenting how datasets and models are analyzed, are created, and would allow third parties, like academic institutions and independent developers, to audit publicly deployed systems for the common good.

Keeping of records and data

We agree with the stated benefits of retaining records of AI datasets and how they were produced. However, forcing disclosure of programming and training methodologies for AI systems could ultimately reduce the incentive of firms to apply AI to certain use-cases, since the algorithmic implementations and training approaches are frequently where companies invest in developing proprietary IP. For example, for some private sector entities, determining the right objective to train against can represent a high percentage of a system’s upfront research and development costs.

To use an analogy, it seems appropriate for regulators to ask a factory to keep detailed records of the raw materials and their sources (the data) it uses to make its products, as well as to keep detailed records of the finished products (the outputs of the AI system), but it would seem like an area for overreach to compel the factory to disclose the precise nature of its production line or machine tooling as this is frequently where the factory invested in the proprietary systems that give it a competitive advantage. Additionally, many issues that the Commission identifies, like fairness and bias, are more ascribable to dataset selection and attributes, rather than the precise algorithms that are used.

Information provision

We recommend standardizing ways of providing information about AI systems. The Commission should seek to clearly label the AI systems similarly to nutrition labels for food products. This ensures information is consumable to the non-technical reader and comprehensive for a technical audience, although language could adapt based on intended audience and environment. Popular approaches to disclosure include the Model Card.[3] OpenAI adopted the Model Card to accompany our recent “GPT-2” language model code release.[4]

[3] Mitchell et al. Model Cards for Model Reporting. 2019. https://arxiv.org/pdf/1810.03993.pdf
[4] https://github.com/openai/gpt-2/blob/master/model_card.md

Robustness and accuracy

In this section, we draw from our submission on the Ethics Guidelines for Trustworthy AI. We also encourage the Commission to standardize qualifications for robustness.

System robustness, as referenced in our Guidelines submission, refers to more than technical safety. In addition to resilience to attack, the system itself should be subject to risk assessments like red-teaming exercises that examine how the system could be repurposed to malicious ends. These assessments should inform plans for risk mitigation, which can be crafted throughout the product development cycle. We should test system accuracy against the environment it is deployed in. Testing should incorporate analyses of potential biases and the results of tests should be logged. However, because AI systems are probabilistic and context-dependent; it is challenging to make many results "reproducible", instead, we encourage the Commission to focus on creating incentives for thoroughly documenting AI system testing and development.

As proposed in our Guidelines submission, we suggest the Commission foster communication channels among developers, between developers and policymakers, and between developers and users to help validate claims about system properties. Initiatives like the Digital Innovation Hubs and testing centers can facilitate communication.

Our notions of what makes a system "robust" or "safe" will necessarily change over time. Therefore, we think the Commission should encourage a culture of spot checking deployed systems. Implementing this is an area of ongoing research. In our research on trustworthy AI, we study how public or private organizations could conduct third party audits of AI systems. Governments could help to define the goals of spot checks for certain regulatory scenarios, third-party actors could develop testing schemes, and industry actors could trial them. This will lead to standardized tests, which will help implementation broadly. Such checks will help ensure that the systems robustness and safety map to societal needs, and are not based on outmoded definitions which lead to unsafe or unrobust behaviours in reality.

It is challenging to develop hard requirements that can "deal with errors or inconsistencies during all life cycle phases" due to the aforementioned context-dependent, socio-technical ways in which AI systems are deployed. Instead, we recommend combining regulations that ensure the system is *developed* in a way that is sensitive to broader safety and robustness needs, and regulations to carry out *spot checks* to ensure a system is functioning as intended.

Human Oversight

Human oversight is an essential part of managing the risks associated with AI development. As a practical step, the Commission could create a task force to determine the best approaches for when and how to conduct this type of intervention, composed of representatives from various member states as well as non-governmental institutions including academia, NGOs, and private

enterprise. Doing so would help hold developers accountable, standardize the process by which claims are verified, and also provide feedback on these “spot checks” to establish a more routine protocol.

E. Addressees

The Ecosystem of Excellence should help guide systems while in development; for example testing facilities can promote robustness and accuracy. This should encompass researchers as well, not just “economic operators”; research can have risk and human impact when deployed, and many researchers develop AI systems by developing prototypes, releasing them into the world, and studying what happens. While many research experiments are covered by institutional controls (e.g., university Institutional Review Boards), the AI community contains many practitioners who are unaffiliated with specific institutions or companies. Oversight tools can help incentivize researchers to develop safe, trustworthy AI systems.

G. Voluntary Labeling

We agree with the benefits of voluntary labeling schemes and encourage standardizing these schemes across a variety of AI systems, then developing an enforceable set of criteria for describing high-risk applications. As explained in the “Defining Risk” section above, AI’s omni-use abilities can lead to a seemingly low risk AI system in a low risk setting adapting to newer, higher risk environments. These low risk systems can be screened, but do not need to be enforced. As explained in the White Paper, Product Liability Directive 85/374/EEC is difficult to apply to AI; proving defects and causality between damage and an AI system is difficult. The difficulty in proving problematic decisions and impact necessitates safeguards in all systems regardless of application. A quality label for low-risk systems should be standardized, with guidance from industry and academia who are developing approaches to labeling and documenting systems.

H. Governance

We believe that a cross-country governance structure can help the Commission address the transnational nature of AI and its associated governance challenges. Such a structure would benefit from a permanent secretariat along with an assembled committee of experts. By having a permanent secretariat, it would be possible to fund and conduct continuous measurement, assessment, and “spot check” activities, which would provide valuable information for EU citizens, elected officials, and the assembled committee of experts. Possible members of the secretariat community could include institutions like the OECD’s AI Policy Observatory, with which the Commission is already collaborating via the Joint Research Center. This governance structure could include permanent members from multiple European countries to reflect both regional and sub-national concerns.

An Ecosystem of Excellence

The ecosystem of excellence and ecosystem of trust should support one another. All proposed actions should support the seven key requirements in the Guidelines for Trustworthy AI and types of requirements by risk.

B. Focusing the Efforts of the Research and Innovation Community

We support testing centers, especially as they can be used to fulfill checklists in the Ethics Guidelines. The Commission should also use these centers to monitor AI progress and help research centers coordinate via creating trans-European benchmarking exercises, where multiple participants contribute to benchmarks to help study AI progress and impact in different areas. These exercises could focus on the sectors where Europe has the potential to become a global champion (industry, health, transport, finance, agrifood value chains, energy/environment, forestry, earth observation and space), kickstarting submission periods to provide a controlled environment to, create standards for, stimulate research, and test AI systems in a given sector.[5] Encouraging technical actors and academia to participate and submit benchmarks can give valuable external information about AI progress and direction. These exercises could also stimulate further research and private sector investment.

C. Skills

Encouraging skills in AI should encompass all relevant interdisciplinary skills and foster skills among EU policymakers and staffers. AI research, development, and policymaking needs interdisciplinary perspectives to more holistically build trustworthy and safe AI. In early 2019, OpenAI published a paper titled *AI Needs Social Scientists* that explains how social scientists are necessary to improve our understanding of the human side of AI alignment.[6] Other examples of skills needed include threat modeling, bias analysis, and economic and labor impact. Skill development should also focus on EU policymakers and staffers to keep them abreast technological changes and to inform AI decisions and policies.

---

[5] For example, the U.S. 'xView' competition in 2018 stimulated research into a certain type of satellite imagery analysis. This could be adapted to Europe by orienting around satellite-based monitoring of agricultural yields.

[6] https://openai.com/blog/ai-safety-needs-social-scientists/

E. Partnership with the Private Sector

We support Innovation Hubs and testing centers, especially as these can be used to facilitate AI evaluation, benchmarking, and standardization. The Commission should use these centers to aggregate information about AI, both building policymaker skills and knowledge, and contributing to efforts to assess AI progress. Testing facilities can each be designated to cater to a specific area of AI; for example, one testing facility may focus on drones and self-driving vehicles while the other may focus on satellite imagery. This would allow different testing sites to develop excellence in different areas of applied AI and strengthen particular regions and sectors.

G. Securing Access to Data and Computing Infrastructures

We strongly support the Commission's proposal on data and computing infrastructures and emphasize two key initiatives:

- **Funding the creation of representative datasets**, and
- **Creating a credit system to make it easy for academics to access subsidized cloud computing services.**

Representative datasets: As discussed earlier, building representative datasets can address bias and fairness concerns. As detailed in the European Strategy for Data, the EU should lead by example in producing data that mitigates fairness concerns.

Cloud credits: The Commission should fund and subsidize academic access to cloud computing services. This is because cloud computing infrastructures (e.g., European cloud companies such as OVH, or American-based ones such as Amazon's Amazon Web Services) let academics access "on-demand" computing services, allowing them to run large-scale scientific experiments without investing in expensive on-premise equipment that quickly depreciates in utility and value. In our research on trustworthy AI development, we identify public benefits to improving academic access to computing power like greater scrutiny of industry claims and open alternatives to commercial systems. Like most AI research organizations, OpenAI predominantly uses cloud computing infrastructures for our computational needs.[7]

[7] In creating such a scheme, the Commission should be wary of the competitive implications, but could mitigate these by creating a standardized Commission-created "cloud credit", which could then be exchanged for resources on any number of cloud computing services. This leaves the company choice up to the users (e.g., academic researchers).

Conclusion

We commend the Commission on its thoughtful approach to excellence and trust in AI and support regulation and funding that leads to beneficial and safe AI. Our comments and recommendations are based on our technical experience and seek to strengthen proposed actions. In accordance with our mission, we are happy to serve as a resource for technical expertise and would be keen to answer any questions our submission raises. We also encourage the Commission to invest in its own technical capacity and in-house expertise to enhance its well-calibrated, effective regulatory actions. OpenAI looks forward to the Commission's continued work on AI.

Submitted by:
Irene Solaiman, Policy Researcher
Bianca Martin, Special Projects Manager
Jack Clark, Policy Director
OpenAI