In this article, the writers examine the current status of the upcoming European Union Artificial Intelligence Act (the “EU AI Act”) in light of the recent three safety recommendations made by Sam Altman, CEO of @OpenAI.

What were the three safety recommendations?

Earlier this month Altman provided three safety recommendations, all of which market participants resonated with. Notably, he specifically left out smaller open-source models from these recommendations.

Firstly, he proposed the formation of a new agency that licenses any effort beyond a certain scale of capabilities. This agency would have the power to revoke such licenses and enforce compliance with safety standards.
Secondly, he suggested creating a set of safety standards, particularly focusing on dangerous capability evaluations. One example previously used is to verify if a model can self-replicate or self-exfiltrate into the wild.

Thirdly, he recommended mandatory independent audits, not only from the company or the agency but from experts who can confirm if the model adheres to the stated safety thresholds and the percentage of performance on various questions.

**What were the latest revisions to the EU AI Act?**

On 9th May 2023 the Committee on the Internal Market and Consumer Protection and Committee on Civil Liberties, Justice and Home Affairs released draft compromise amendments on the draft report. These have been proposed ahead of the vote by the triforce – European Commission, European Parliament and the European Council – between 12-15 June. Continued revisions to the EU AI Act are expected to be made as the technology continues its unprecedented adoption by the modern economy. Fears over its capabilities primarily drive the pace of adoption, which legislators then take ownership of to prevent harms being caused to consumers.
### Safety recommendations versus the latest text of the EU AI Act

<table>
<thead>
<tr>
<th>Safety Recommendation</th>
<th>EU AI Act</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Agency/Office</strong></td>
<td></td>
</tr>
<tr>
<td>Recital (61b)</td>
<td>AI office must be represented in efforts to enhance multistakeholder governance.</td>
</tr>
<tr>
<td>Recital (61b)</td>
<td>AI office must be consulted by the European Commission in the preparation of the standardisation request in order to collect relevant expertise.</td>
</tr>
<tr>
<td>Article 53 (5a)</td>
<td>Establishing authorities shall inform the AI Office of the establishment of a sandbox and may ask for support and guidance.</td>
</tr>
<tr>
<td>Recital (76)</td>
<td>The AI Office should have legal personality, should act in full independence, should be responsible for a number of advisory and coordination tasks, including issuing opinions, recommendations, advice or guidance on matters related to the implementation of the EU AI Act and should be adequately funded and staffed.</td>
</tr>
<tr>
<td>Article 56</td>
<td>Establishment of the AI Office as an independent body of the European Union. It has a legal personality.</td>
</tr>
<tr>
<td>Article 56 (b)</td>
<td>Prescribed tasks assigned to the AI Office, including supporting, advising and cooperating with EU Member States; monitoring and ensuring the effective and consistent application of the EU AI Act; coordinating joint investigations; and serving as a mediator in certain discussions.</td>
</tr>
<tr>
<td>Article 56 (c)</td>
<td>Prescribed accountability, independence and transparency requirements for the AI Office, including remaining accountable to the European Parliament and European Council.</td>
</tr>
<tr>
<td><strong>Standards</strong></td>
<td></td>
</tr>
<tr>
<td>Recital (61)</td>
<td>Imposition of standards as a policy tool for ensuring conformity with the requirements of the EU AI Act.</td>
</tr>
<tr>
<td>Article 40 (1a)</td>
<td>The European Commission issues standardisation requests covering all of the requirements of the EU AI Act.</td>
</tr>
<tr>
<td><strong>Independent Audits</strong></td>
<td></td>
</tr>
<tr>
<td>Recital (60h)</td>
<td>Third-party auditing measures remain under development.</td>
</tr>
<tr>
<td>Recital (60h)</td>
<td>New ways to assess fundamental models that fulfil in part the objective of auditing (such as model evaluation, red-teaming or machine learning verification and validation techniques) are under development.</td>
</tr>
</tbody>
</table>
Current state of the EU AI Act

How do the safety recommendations align with the EU AI Act?

Evidently, most of Altman’s recommendations are taken into consideration by EU legislators in the latest draft. While the alignment isn’t on a like-for-like basis, the ‘spirit’ and ‘principle’ of what has been incorporated into the legislative provisions. For example, Altman recommends that third-party independent audits are mandatory, whereas the EU AI Act sets down a requirement to conduct a third-party independent audit while also considering standards to assist certified bodies in undertaking this. Promising for industry participants to note that these recommendations have already been considered or are in the process of being considered to tackle the risks arising from AI systems.

Best practices for developing safe and trustworthy AI

We have identified three best practices that leading AI firms can use to build safe and trustworthy AI systems:

1. Have clear objectives during the development process

One of the main reasons that AI systems cause concerns is because of an absence of set objectives, aligning stakeholders across the entire value chain. It is imperative to align these stakeholders from the start of the development process create a harmonised direction of travel.

This helps to ensure that all stakeholders acknowledge and understand what is trying to be achieved and why. For example, leading AI firms may often start development processes with pure commercial objectives without taking the full suite of safety elements into consideration.

In essence, aligning on the vision, design, and trade-offs for developing AI systems improves transparency across the business while uncovering and reducing risks.
2. Build a robust technology development plan

Builds are about more than just introducing new technologies. For successful projects, we are of the view that three key elements matter most:

- **Be intentional** when considering the safety elements (for example, focus on the influencing factors, not just the end state).
- Implement a **cooperative** mindset that intersects technology and business goals (for instance, for the user, by the user—involves the business at all times and in all phases of the process).
- Drive forward with **comprehensive** and transparent execution tailored to the realities of modern AI systems (for example, apply a modular approach to plug and play).

Nonetheless, all technology builds experience setbacks. Employees often misunderstand or do not appreciate the nuances of the ethical elements, and these can result in unforeseen risks.

To secure early support during the pilot phase and to help minimise risks, an AI firm could run existing rule- and scenario-based risk scenarios in parallel with safety-based scenarios to build confidence among stakeholders.

To further build adoption and reduce risk, the AI firm might choose projects that can leverage existing safety standards (those that employees are already cognisant of) and integrate the new components in consecutive fashion.

Ideally, AI firms will start with the low-hanging fruit—projects that offer significant potential rewards with manageable risk.

3. Optimise the model development process

To build safe and secure AI systems for use by the business functions, functional teams need to conduct the following:

- **Enlarge understanding** to work closely with the ethics team in the model development and validation process. Functional teams should educate personal about the necessary safety elements during the development process.
Current state of the EU AI Act

- Direct validation standards, policies, and frameworks to take on the specific safety risks linked to AI systems, including bias detection and explainability.
- Define specific performance and monitoring requirements, including out-of-time testing, and when to recalibrate ML models.

Best practices for bringing safety to AI systems

In the mission to build safe AI systems, AI firms have traditionally been one step ahead of regulatory standards from an operational standpoint. At present, they have a chance to positive influence the industry. Safety standards enshrined under the EU AI Act promise to improve the AI ecosystem dramatically by reducing harms caused to end users. For some AI firms, the development requires investing significant time and resources. To remain fully compliant with the safety standards, AI firms will need to remain up to date with the safety standards, align the development teams, and acquire senior management buy-in. A reasonable challenge, but well worth the effort given the high stakes, as shown by the recent senate committee.

How much will new US and EU rules and safety recommendations impact innovation and economic growth?

The recent introduction of the AIA in Europe and possible new regulations in the United States have raised concerns about their potential impact on innovation and economic growth.

According to a report by Benjamin Mueller 2021-aia-costs.pdf, “the AIA will cost the European economy €31 billion over the next five years and reduce AI investments by almost 20%. This could have a significant impact on small and medium-sized enterprises (SMEs) that deploy high-risk AI systems, as they may incur compliance costs of up to €400,000, causing profits to decline by 40%.”

To quote President Biden on Twitter: “When it comes to AI, we must both support responsible innovation and ensure appropriate guardrails to protect folks’ rights and safety. Our Administration is committed to that balance, from addressing bias in algorithms – to protecting privacy and combating disinformation.”
Current state of the EU AI Act

The National Telecommunications and Information Administration (NTIA) is seeking public input to boost AI accountability. “The insights gathered through this Request for Comment (RFC) will inform the Biden Administration’s ongoing work to ensure a cohesive and comprehensive federal government approach to AI-related risks and opportunities.” As of May 19, 2023, 464 comments have been received, with comments due on or before June 12, 2023.

Overall, it remains unclear how much these new rules and potential new rules, and 3 safety recommendations: (1) Agency/Office/License, 2) Standards, 3) Independent Audits made by Sam Altman, will or could impact innovation and economic growth in Europe and the United States. It is important for policymakers to carefully consider the potential costs and benefits of these regulations before implementing them.

Independent audits can be leveraged to maximize responsible innovation and ensure appropriate guardrails?

How AI & Partners can help

We can help you start assessing your AI systems using recognised metrics ahead of the expected changes brought about by the EU AI Act. Our leading practice is geared towards helping you identify, design, and implement appropriate metrics for your assessments.

Website: https://www.ai-and-partners.com/

@AI_and_Partners