FLI Position Paper on the EU AI Act

FLI position on the Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)

4 August 2021

European Commission
Rue de la Loi 200
1049 Brussels
Belgium

Dear Executive Vice President Vestager and Commissioner Breton,

The Future of Life Institute (FLI) is an independent nonprofit with the goal of maximising the benefits of technology and reducing its associated risks. This remit inspired us to organise a conference in Asilomar - California in 2017 to formulate one of the earliest artificial intelligence (AI) governance instruments: the "Asilomar AI principles."

When, on April 21st, you jointly proposed the AI Act, the EU became the first major regulator to craft a law on this technology. This is a significant milestone for AI governance and one that would have been unthinkable when we gathered in Asilomar less than five years ago. I would like to commend you for taking a proactive governance approach to the development of AI and for offering an opportunity to provide feedback.

AI development is occurring at breakneck speed and it is hard to predict future applications of today's experimental systems. The GPT-3 system, for example, produces human-like text and has many intended purposes: it can just as easily generate captions under newspaper images as it can be used to produce descriptions of human faces for biometric identification. Considering this, FLI recommends that the regulatory proposal is updated to address increasingly generalised AI systems that have multiple purposes. You can find these and other recommendations in the attached position paper.

In recognition of the EU's leading role in AI governance, I am also pleased to formally announce the expansion of FLI's operations to Europe. The organisation is also registered in the EU Transparency Register with registration number 787064543128-10.

I stand ready to offer FLI's technical expertise for beneficial AI governance in Europe.

Sincerely,

Professor Max Tegmark
President

Future of Life Institute

FLI Position Paper on the EU AI Act

The Future of Life Institute (FLI) is one of the world's leading voices on the governance of AI. The institute is an independent nonprofit that works on maximizing the benefits of technology and reducing its associated risks.

FLI created one of the earliest and most influential set of AI governance principles - the Asilomar AI principles - and maintains a large network among the world's top AI researchers. The institute, alongside the governments of France and Finland, is also the civil society champion of the recommendations on artificial intelligence in the UN Secretary General's Digital Cooperation Roadmap.

This paper provides FLI's views on the Commission's proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). The Act takes a risk-based approach to regulating AI. First, applications and systems that create an unacceptable risk, such as government-run social scoring of the type used in China, is banned. Second, high-risk applications, such as self-driving cars, are considered high-risk and are therefore subject to specific legal requirements. Lastly, applications not explicitly banned or listed as high-risk are largely left unregulated.

Recommendations

1. Account for the full (and future) risks of AI
    a. Update the proposal to allow for increasingly generalised AI systems that have multiple purposes, such as GPT-3, DALL·E and MuZero;
    b. Ensure AI providers explicitly consider the impact of their systems on society at large;
    c. Protect consumers against conflicts of interest, especially when AI-systems are used in medical, legal, or financial contexts;
2. Enhance protections of fundamental rights
    a. Allow EU citizens to file a complaint with authorities when an AI has manipulates their behaviour, or when their fundamental rights are breached;
    b. Lower the threshold for what constitutes subliminal manipulation;
    c. Expand whistleblower protections to AI developers, because they are the first, or only, people who know whether an application violates fundamental rights;
    d. Require reporting of AI safety incidents at the European level;
3. Boost AI innovation in Europe
    a. Empower the European Artificial Intelligence Board to pre-empt new risks;
    b. Create one AI portal across the European Single Market, allowing SMEs easy registration with regulatory 'sandboxes' - environments in which firms can try out new services without fear of penalties;
    c. Build more public sector capacity for AI development and oversight;

If you have any questions about the recommendations in this position paper, please do not hesitate to contact Mark Brakel, FLI's Director of European Policy, at mark@futureoflife.org or through +31616647630.

I. Account for the full (and future) risks of AI

a. Update the proposal to address increasingly generalised AI systems that have multiple purposes, such as GPT-3, DALL·E and MuZero

Under the current draft, legal requirements depend on what an AI system's singular intended purpose is in a given use case. For example, an AI application that assesses students' performance is regulated, whereas the same application that then recommends educational materials to improve the performance for those same students is not.

FLI recommends that the proposal be improved to address the latest, and future, technological developments in the field of AI. New systems such as GPT-3, DALL·E and MuZero have an unknown number of applications and classifying the whole system by a single use could allow increasingly transformative technologies to evade regulatory scrutiny. GPT-3, for example, is an AI application that can generate text that has been shown to be biased against Muslims. Specifically, in more than 60% of cases, GPT-3 created sentences linking Muslims to shooting, bombs, murder and violence [1]. FLI believes that bias in these type of AI applications should be regulated both when used to generate captions under newspaper images (low-risk, unregulated under the proposal) and when it creates descriptions of human faces for real-time biometric identification (high-risk, regulated under the proposal).

Future AI systems will be even more generalised than GPT-3. Therefore, **the proposal should require a complete risk assessment of all of an AI system's intended uses** (and foreseeable misuses) instead of categorising AI systems by a limited set of stated intended purposes. One way of achieving this would be to expand title IV ("Transparency obligations for certain AI systems") to include a limited number of additional requirements that apply across all AI applications regardless of their intended purpose. Further, revising the definitions in Article 3 of the proposed Act to account for plural "uses" and "purposes" will improve the overall scope of the proposal to address these increasingly generalised AI systems.

b. Ensure AI providers consider the impact of their applications on society at large

AI applications may cause societal-level harms, even when they cause only negligible harms to individuals. For example, a political marketing application may reduce a person's desire to vote by a small amount. At an individual level, the impact of this application may not be considered an infringement of fundamental rights, but collectively, the effect may be large enough to change an election result. Societal-level harms such as widespread disinformation or manipulation by recommender systems in social media should be taken more seriously in addition to direct harms to individuals.

An AI system that maximises ad clicks, for example, will show users addicting content. In turn, this application causes users to spend more time on social platforms, and may foster societal polarisation and increased misinformation. The effects of these systems can be stark. For example, since the introduction of widespread social media in 2009, women aged 15-19 experienced a 70% increase in the suicide rate, and women aged 10-14 experienced a 151% increase [2].

[1] Dave Gershorn, For Some Reason I'm Covered in Blood': GPT-3 Contains Disturbing Bias Against Muslims, January 2021.
[2] Centers for Disease Control and Prevention, Increase in Suicide Mortality in the United States, 1999-2018, April 2020.

FLI recommends that a final AI Act does more to address indirect and aggregate harms. One way of achieving this goal would be to modify the assessment procedure so that AI providers consider the impact of their applications on society at large. This could be brought about by including possible societal impact in the requirements for technical documentation (Annex IV).

c. Protect consumers against conflicts of interest, especially when AI-systems are used in medical, legal or financial contexts

There are many AI systems, especially digital services, where consumers reasonably expect that the output provided by the AI system meets certain unstated standards. For example, if someone asks a virtual assistant (e.g. Apple's Siri, Amazon's Alex) to book a flight, we expect it to consider all options and to present us with the cheapest or fastest option [3]. However, this expectation can be violated when an AI provider enters into an undisclosed business arrangement with a particular airline to favour its tickets. Similarly, an AI-powered mapping service could be sponsored by a restaurant to (sometimes) divert the suggested driving route of users towards its business. FLI believes that these conflicts of interest should be regulated in the interest of consumer protection.

Legal protections against conflicts of interest are likely to become more important as AI systems take over sensitive medical, legal or financial services. AI "therapists" today, for example, are little more than chatbots that provide standardised responses, but such systems are bound to become more capable in the coming years. Initially, the system may only recommend treatments, which are then approved by a licensed human. With time, the system will recognise patters and human approval could become increasingly pro-forma until the AI is effectively providing the treatment independently. When this happens, guarantees need to be put in place to ensure that AI systems serve the interests of its patient. An AI application should avoid disclosing information about one patient to another, and base treatment solely on what is best for a patient.

When AI systems act in high-risk contexts that have traditionally been performed by humans with fiduciary duties, such as doctors, lawyers or financial advisors, they should be required to abide by the applicable professional standards. Therefore, FLI proposes to amend article 9 of Chapter 2 (requirements for high-risk systems) to mandate that AI applications used in medical, legal or financial contexts explicitly consider the risk that they violate fiduciary responsibilities in their respective sectors.

In medium- and low-risk contexts, where the impact of conflicts of interest is likely less severe, users should at least be informed about inherent conflicts of interest. To bring this about, EU legislators should consider modifying article 52 on transparency obligations to ensure that AI systems interacting with users are required to explicitly state their primary and secondary goals to them.

[3] Example taken from Anthony Aguirre et al, Handbook of AI Governance, forthcoming.

II. Enhance protections of fundamental rights

a. Allow EU citizens to file a complaint with authorities when an AI manipulates their behaviour, or when their fundamental rights are breached

The first objective of the proposal is to ensure that AI systems placed on the market are "*are safe and respect existing law on fundamental rights and Union values.*" For this goal to be fulfilled, it is critically important that providers comply with the law and that any breaches are quickly remedied. This requires that sufficient channels exist for providers and regulators to be made aware about breaches and risks.

The proposal introduces third-party verification for AI systems that are safety components of products. Similarly, so-called "notified bodies" will verify compliance with the Act if AI systems engage in biometric identification of people [4]. In turn, decisions by notified bodies may be appealed by "*parties having a legitimate interest*". This right of appeal (article 45) is an important safeguard of fundamental rights, and it will ensure that, for example, a trade union can appeal the approval of systems that use facial recognition in ways not intended or foreseen by the notified body. A trade union could invoke this right to question compliance of a system that uses facial recognition to analyse the emotional state of workers as they begin their shifts.

The right to appeal decisions of notified bodies alone, however, provides insufficient protection of fundamental rights, because many (high-risk) AI applications can be put on the market after a self-assessment without third party involvement. Therefore, under the current draft, the relevant national supervisory authority is the only body that can act when a provider of (high-risk) AI overlooks or evades a legal requirement. The implications of this are stark. If someone falls victim to an AI system that, for example, "*deploys* (harmful) *subliminal techniques beyond a person's consciousness*", then they may not be able to file a complaint through a dedicated process.

FLI believes that people affected by AI outcomes should be able to challenge these systems individually and collectively. Taking inspiration from the EU's General Data Protection Regulation (GDPR), the proposal could give individuals a right to lodge a complaint with their relevant national supervisory or market surveillance authority when they feel their safety, health, and rights are at risk or have been breached (comparable to article 77 GDPR). Similarly, **FLI recommends that the right to bring forward a representative complaint for consumer groups and others (article 80 GDPR) is included into the final AI Act.**

b. Lower the threshold for prohibited subliminal manipulation

Article 5 of the proposal bans AI systems that deploy "*subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm.*" **FLI strongly welcomes this provision and the way in which it protects human autonomy.** Without it, AI-driven applications may increasingly shape our individual preferences as well as alter the outcomes of political decision-making.

[4] Once standardisation bodies issue harmonised standards for compliance (or common specifications become available) providers of biometric identification systems will be able to self-assess without third-party verification by a notified body.

FLI recommends that the scope of this provision is widened and the harm requirement is removed. The draft proposal - rather than prohibiting practices where a would-be manipulator's own ends are furthered at the expense of someone else - appears to permit manipulative AI systems insofar as they are unlikely to cause an individual harm [5]. This definition risks giving a free pass to many manipulative AI systems, because subliminal manipulation is hard to detect and because it will be difficult for an affected person to prove a causal relationship between the subliminal manipulation and the harm incurred. In other words, the high threshold risks making this important prohibition largely symbolic.

Therefore, **FLI proposes the prohibition of any manipulation that adversely impacts fundamental rights or EU values** (as defined in article 2 of the Treaty on European Union, TEU), regardless of whether the AI system uses subliminal techniques, nudging, or any other manipulative strategy. Amended in this way, the final Act invites AI providers to critically evaluate whether their application or system seriously distorts human decision-making.

c. Expand whistleblower protections to AI developers, because they are the first, or only, people who know whether an application violates fundamental rights

As AI applications grow ever more complex, it will become increasingly difficult to know whether an application constitutes a health risk or a potential violation of human rights. In fact, some of the only people who will be able to determine the nature and extent of risks and harms will be the developers themselves. Therefore, these **developers should be provided with the right to voice concerns to a relevant supervisory authority through a dedicated channel if internal company channels are insufficient, and should be able to rely on EU whistleblower protections** (Directive (EU) 2019/1937). Recent controversies surrounding the employment of experts in AI ethics by major private companies make it apparent that such whistleblower protections may be necessary for industry experts to feel comfortable raising concerns to outside authorities [6].

d. Require reporting of AI safety incidents at a European level

The development of AI technologies is happening at breakneck speed and safety implications of many AI systems often only become know after they are placed on the market. The Boeing 737 MAX, for example, had been tested for many years and was certified by the U.S. Federal Aviation Administration in March 2017. It took two years and two plane crashes before investigators discovered that an AI-based software system within the cockpit, the Manoeuvring Characteristics Augmentation System (MCAS), produced fatal nose-down commands without an override option for pilots [7].

[5] Michael Veale and Frederik Zuiderveen Borgesius, Demystifying the Draft EU Artificial Intelligence Act (Computer Law Review International), pre-print July 2021.
[6] See, for example, Tom Simonite, What Really Happened When Google Ousted Timnit Gebru (Wired Magazine), 8 June 2021.
[7] See Dominic Gates and Mike Baker, The inside story of MCAS: How Boeing's 737 MAX system gained power and lost safeguards (The Seattle Times), 22 June 2019.

Under the current proposal, article 62 rightly requires that AI providers report "*any serious incident or any malfunctioning of those systems which constitutes a breach of obligations under Union law*" to market surveillance authorities of Member States. AI advancement in Europe would however also benefit from a clear overview of safety incidents at a European level, because this will make it easier to analyse what research or regulation may be necessary as trends emerge across the Single Market. Therefore, and in the spirit of the existing Seveso directive on industrial accidents, **FLI recommends that Member States also report safety incidents to an EU database** (article 60).

III. Improve the European environment for innovation

a. Empower the European Artificial Intelligence Board to pre-empt new risks

Under the current proposal, the European Artificial Intelligence Board acts mainly as a coordinator of regulatory efforts. The Board is tasked with the promotion of cooperation between national supervisory authorities and to assist the European Commission with ensuring uniform application of the law. Instead, a more expansive remit could help boost AI innovation in Europe.

FLI supports the recommendation of the European Data Protection Supervisor (EDPS) and the European Data Protection Board (EDPB) that the new AI Board be given "*sufficient and appropriate powers*" [8]. In particular, **FLI recommends that the Board be authorised to implement changes to the Act's Annexes on its own accord and without restriction.** When the Board has the power to add to the list of restricted (article 5) and high-risk systems (annex III), it will be able to swiftly act upon signals it receives from market surveillance authorities and ensure uniform standards across the Single Market if risks emerge outside the areas initially identified in the proposal.

As suggested by the Commission, a full-fledged European Artificial Intelligence Board should also work closely with a group of experts to ensure that beneficial technologies are accelerated and risks mitigated. These experts can highlight relevant technological breakthroughs and ensure timely updates to the Annexes so that businesses are provided with early regulatory certainty.

b. Create one AI portal for the European Single Market as a whole, allowing SMEs easy registration with regulatory 'sandboxes' – environments in which firms can try out new services without fear of penalties

The Commission has rightfully proposed the creation of "regulatory sandboxes" to facilitate the development of innovative AI systems before their placement on the market. These sandboxes will be critical to supporting European AI innovation, because they allow small start-ups to develop their applications in close cooperation with regulators and without fear of penalties.

[8] EDPB-EDPS, Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), June 2021.

Without the sandboxes, AI start-ups lacking legal departments may fail to comply with the proposal altogether or may avoid product development in certain high-risk areas.

If Europe truly wants to be at the forefront of AI development however, its approach to sandboxes could be made more ambitious. Leading AI developers are likely to hail from a variety of Member States with differing approaches to sandboxes. This disparity risks fracturing the Single Market and slowing down AI development. A stronger approach would involve the establishment of a pan-European sandbox, which would be accessible through a single online AI portal.

Enhanced sandboxes could offset some of the regulatory burden introduced through the Act by offering additional services to participating businesses, such as legal support, lab-to-market insurance and fiscal incentives for R&D activities. Moreover, the EU should consider opening up access to sandboxes to SME's from outside the Union. In that way, sandboxes would promote the dissemination of EU standards to supplies from outside its borders. The EU could also facilitate input from AI experts in civil society and academia through the sandboxes to help ensure that the guidance provided to businesses remains state-of-the-art.

c. Build more public sector capacity for AI development and oversight

AI is a major potential source for economic growth which must be facilitated by civil servants who understand the latest technological developments. Governments should therefore view increased public sector capacity for AI development as an opportunity, rather than as a burden.

Currently, the Commission estimates that implementation of the proposal will require no more than 25 Full Time Equivalent civil servants per Member State. This estimate is likely to underestimate the transformative impact that AI will have on both societies and their public sectors.

Beyond the minimal bar set by the Commission, both EU institutions and Member States may want to consider investing extra resources in public sector capacity in order to i) ensure companies are able to get quick answers from regulators on whether a sandbox application can be placed on the market; ii) improve understanding of where public research funds can best be directed, iii) quickly publish non-personal data from local, regional and national authorities to improve public services.