THE FUTURE SOCIETY

Proposal for a regulation - "Artificial intelligence – ethical and legal requirements"[1]

Trust in Excellence & Excellence in Trust

Contribution by The Future Society, August 2021

The Future Society (TFS) is a global nonprofit advancing the responsible adoption of Artificial Intelligence (AI) for the benefit of humanity. With a network of policy researchers and practitioners in the EU, the US and all over the world, we build understanding of AI and its impact, we build bridges between relevant constituents, and we build innovative solutions to help communities and people all over the world enjoy the benefits of AI and avoid its risks. Incorporated since 2016 and funded through a diverse mix of donors and a broad community, we provide an independent and nuanced perspective on the governance of AI.

TFS contributes to the governance of AI on multiple fronts. Recently, we have:

- *supported the OECD's work as part of the expert group that drafted its OECD AI Principles,[2] through commissioned research on AI Value Chains[3] and now as an active element of its AI Policy Observatory;*
- *undertaken research & strategic advisory work for the Global Partnership on AI, after winning several competitive tenders and as selected experts in various working groups and committees (Responsible AI, AI for Pandemic Response, Data Governance, AI for Climate Change, etc.);[4]*
- *co-founded the Athens Roundtable on AI and the Rule of Law, bringing senior decision-makers from the US and the EU together to co-shape the governance of AI;[5]*
- *advised governments and development IGOs on matters related to AI governance, as with our recent work developing Rwanda's National AI Policy in partnership with the Rwandan government and GIZ.[6]*

TFS is a nonprofit registered in both Boston, United States and Tallinn, Estonia. We have joined the Transparency Register (ID# 473310732515-30) in 2018 following our strong conviction that the EU institutions have a leading role to play in the governance of AI. We remain at your disposal if you have any questions related to our work: info@thefuturesociety.org.

[1] Initiative details - Artificial intelligence – ethical and legal requirements
[2] OECD AI Group of Experts (AIGO)
[3] Mapping the AI Value Chain
[4] The Global Partnership on AI (GPAI) on Responsible AI and AI in Pandemic Response
[5] The Athens Roundtable on Artificial Intelligence and the Rule of Law 2020
[6] The Development of Rwanda's National Artificial Intelligence Policy

Trust in Excellence & Excellence in Trust

Summary of recommendations

A. Investing in smart governance capabilities for AI and the Digital Single Market

> Stronger Testing & Experimentation Facilities and other governance mechanisms building cutting-edge expertise for the public sector

- Invest in properly staffing and resourcing the mechanisms for the governance of AI;
- Establish clear and coherent mandates and maintain accountability;
- Prevent undue political interference, but welcome public consultations related to consumers preferences and fundamental rights.

> A more ambitious regulatory sandbox system

- Offer an EU-wide, one-stop, SME-friendly sandbox solution;
- Offer additional pre-deployment, compliance & R&D support services to make the sandbox solution more attractive;
- Open the sandbox to foreign entrepreneurs interested in developing “Tech fit for EU”.

B. Ensure governance is adaptive and responsive to macro-trends

> A future-ready AI Act curtailing harmful macro-trends

- Provide for the specification of multiple purposes and unintended purposes for each AI system;
- Require limited but informative horizontal safeguards for specific technologies based on objective features of the system itself rather than the stated intended purposes (e.g. number of parameters, physical forces involved, ...);
- Require providers to discuss the macro-level ethical and social implications of this type of system if deployed at scale, similar to academic publications guidelines.

> A more effective governance response to technological trends and their implications

- Ensure information flows between national and European levels of governance and that this information is ready for EU-wide analysis;
- Invest in the capacity to systematically compile at the EU-level and analyse incident reports from all Member States;
- Ensure the expert group supporting the European AI Board include a significant share of members from consumer organisations and civil society;
- To limit regulatory uncertainty, reduce the time between a concerning finding from the European AI Board and the Commission’s decision to adapt the regulatory framework or the justification for not adapting it.

C. Avoid a "lemons market" by fostering trustworthy market dynamics

> Protect long-term market profitability from the information asymmetry's economic implications

- Preserve the AI Act's current allocation of obligations to the providers with regards to the requirements for which they have most information and effective control; prevent undue shift of obligations onto users;
- To the extent it can be achieved efficiently, facilitate the flow of technical information and understanding from the provider to its users.

> Empower citizens to defend their fundamental rights, health and safety more directly

- Require users to report serious incidents and malfunctioning to the relevant authorities, not only to the provider of the defective AI system;
- Establish a direct complaint procedure and extend individuals' right to not be subjected to non-compliant systems;
- Preserve the possibility for parties having a legitimate interest to appeal notified bodies' decisions.

> Provide ex-ante measures to foster more trust in the market

- Directly connect the AI Act to relevant measures in Directive (EU) 2019/1937 protecting whistleblowers;
- Require disclosure of the AI system's loyalty to users and factor in the importance of the loyalty of an AI system in the AI Act for consumers trust;
- Consider additional requirements for providers to foster trust as efficiently as possible.

We explain our full reasoning behind these recommendations below.

Background

TFS participated in the European Commission's public consultation on the White Paper on AI in 2020. A summary of that contribution is available here: the overall recommendation was to develop world class auditing capabilities and Testing & Experimentation Facilities. In doing so, we highlighted the importance of:

- Evidence-based design of these facilities and associated policy instruments;
- Leveraging both ex-ante and ex-post compliance tools;
- Opening access internationally to disseminate the EU's quality standards abroad;
- Ensuring SMEs, local governments, researchers, entrepreneurs, … have easier access to these facilities;
- Making the testing & experimentation facilities "learning organisations", adaptive to change, including regular landscape reviews;
- Investing in research, innovation & capacity-building for testing and auditing technologies.

We are glad and grateful that many of these recommendations have been integrated in the AI Act or in the associated debate. We feel the European Commission is setting itself up for success by considering adaptive, forward-looking instruments like the European AI Board and

flexible tools like the sandboxes. Building on the evolution of the debate surrounding the AI Act in the past year, we therefore make additional recommendations to further improve on these and other aspects.

A. Investing in smart governance capabilities for AI and the Digital Single Market

In the AI Act, the European Commission has proposed the development of interesting governance capabilities, such as the European AI Board and the regulatory sandboxes. We welcome this ambition: ensuring the balance of excellence and trust in the pervasive, fast-evolving field of AI will require a strong & effective governance capacity. We suggest that the European Commission continues to build the EU's governance capacity in this field through the careful design of Testing & Experimentation Facilities (TEFs) and the sandboxes.

Indeed, we believe that successful TEFs and sandboxes could be the best stimulators for both innovation and trustworthiness in the EU. Properly staffing and funding these new mechanisms should therefore be a priority investment, not only for the EU's leadership in trustworthy AI, but also for the Digital Single Market as a whole. Given the technology and products are often similar across Member States, we suggest a tightly interwoven network of national labs and EU centers, which would build European authorities' capacity to govern AI and avoid unnecessary fragmentation of the market. We elaborate on the design and development of TEFs in our previous position paper.[7] As explained there, these smart governance capabilities can be used as a way to disseminate the EU's trustworthy AI approach. The arguments equally apply to the sandbox system: they should be designed as a way to attract innovators from all over the world -from EU SMEs to foreign entrepreneurs and researchers- and therefore should facilitate compliance through administrative assistance and other innovation-friendly support services and benefits.

Moreover, there are positive externalities from building expertise and experience within the civil service at the EU-level, through TEFs and sandboxes and beyond. There will be greater confidence of citizens and industry in the efficient enforcement of the regulation, greater accountability and responsiveness to various stakeholders and greater ability to advise other public agencies in the EU about the deployment of AI and data-sharing for public services without conflict of interests. Attracting the experts in these fields will be costly given the current competition for talent within industry, but the independence of this expertise provides qualitative gains for EU governance. This investment of taxpayer money will be significant, and it is therefore crucial to have clear accountability mechanisms in place, as we described in our previous position paper.[8]

[7] The Future Society - Experimentation, testing & audit as a cornerstone for trust and excellence
[8] Ibid.

Recommendations:

> Stronger Testing & Experimentation Facilities and other governance mechanisms building cutting-edge expertise for the public sector

- Invest in properly staffing and resourcing the mechanisms for the governance of AI;
- Establish clear and coherent mandates and maintain accountability;
- Prevent undue political interference, but welcome public consultations related to consumers preferences and fundamental rights.

> A more ambitious regulatory sandbox system

- Offer an EU-wide, one-stop, SME-friendly sandbox solution;
- Offer additional pre-deployment, compliance & R&D support services to make the sandbox solution more attractive;
- Open the sandbox to foreign entrepreneurs interested in developing “Tech fit for EU”.

B. Ensure governance is adaptive and responsive to macro-trends

As we have witnessed over the past couple of years, AI as a technology is evolving fast. For example, Deepfake or chatbot technologies have moved from quirky research projects to widespread, consumer-ready applications. Moreover, while the implications of AI technology on individuals are increasingly understood (e.g. a recommender system might over time generate addiction or shifts in beliefs), the aggregation of these individual effects at the macro level is still not properly addressed (e.g. spread of conspiracy theories, relation between suicide and social media usage, flash crashes, …). Yet, these macro-level effects can significantly impede the health, safety and fundamental rights of tens of millions of citizens. For example, a European small business owner might be profiled by recommender systems as “too small” for being shown online advertisements or news about public tenders as often. At the individual level, this slight disadvantage might not even be admitted as a loss of opportunity. However, given how widespread various AI-based recommender systems are and compared to when small business owners were reading the same magazines as multinational companies, this would generate the macro effect of reducing access to public tenders to SMEs EU-wide. This would have macroeconomic implications, such as a reduced ability for SMEs to scale up through big public contracts.

The AI Act, as it stands today, cannot deal with the speed of change in the technological landscape nor with the technology’s macro-implications: its focus on the intended purpose of the AI system inherently constrains governance mostly to the immediate, micro-level interactions between the system and its environment. We welcome the inclusion of AI system life-cycle considerations when addressing risk management (in Article 9) and of interactions with other systems when addressing robustness requirements (in Article 15). However, these requirements apply to a narrow subset of all AI systems’ intended purposes and won’t be sufficient to prevent undesirable macro-trends arising from the technology’s overall progress rather than a specific system’s intended purpose. To address these, we suggest improvements to the substance of the AI Act and to the associated governance system.

First, to ensure the AI Act remains relevant in the future, it should explicitly consider that a single system can be used in many different ways. The Act should enable the specification of multiple purposes (intended and unintended) to be assessed for compliance and around which it should apply safeguards. Moreover, as it is not a single system that gives rise to macro trends but the deployments of multiple systems (as in the example of recommender systems above), these safeguards should not be purpose-specific but rather "technological type"-specific. In order to limit the regulatory burden, the scope of application could be limited by features of the system itself (number of parameters, number of users, revenue size of the provider, physical forces involved or extent of mental affect, …) and their requirements could take the form of a registration and provision of information. The AI Act should also require providers of high-risk AI systems and encourage other providers to disclose and discuss the macro-trends that could follow from deploying their and similar systems at scale - similar to the growing number of publications in the field that require authors to discuss the potential negative ethical and societal consequences of their research.

Second, the governance system should also be adapted to better prevent and mitigate macro-trends. A key strength of the EU is its strong, cohesive network of 27 national governance systems. By ensuring good coordination and exchange of information between the national and EU level, this enables the detection of macro-trend patterns more easily (e.g. if 4-5 Member States authorities face similar challenge). To guarantee the EU can reap this "low hanging fruit" for efficient governance, we recommend an aggregation of information, notably of incident reports, at the EU-level and investment in the capacity to analyse this information with the aim to detect macro-trends and recommend potential responses. This analysis role could naturally fall within the European AI Board (as already hinted at in Articles 56(2)(b) and 58), especially if informed by an associated expert group that include consumer associations and civil society. We consider however that the envisaged European AI Board will not have sufficient powers to speedily address issues raised. To limit the regulatory uncertainty triggered by a debate or a red flag raised by the Board, we recommend that the European AI Board's recommendations be strong enough to automatically launch a discussion within the European Commission about whether and how to amend the relevant Act's Annex(es) accordingly, and that rejection of the European AI Board's recommendation be quickly justified by a response from the European Commission.

Recommendations:

> A future-ready AI Act curtailing harmful macro-trends

- Provide for the specification of multiple purposes and unintended purposes for each AI system;
- Require limited but informative horizontal safeguards for specific technologies based on objective features of the system itself rather than the stated intended purposes (e.g. number of parameters, physical forces involved, ...);
- Require providers to discuss the macro-level ethical and social implications of this type of system if deployed at scale, similar to academic publications guidelines.

> A more effective governance response to technological trends and their implications

- Ensure information flows between national and European levels of governance and that this information is ready for EU-wide analysis;
- Invest in the capacity to systematically compile at the EU-level and analyse incident reports from all Member States;
- Ensure the expert group supporting the European AI Board include a significant share of members from consumer organisations and civil society;
- To limit regulatory uncertainty, reduce the time between a concerning finding from the European AI Board and the Commission's decision to adapt the regulatory framework or the justification for not adapting it.

C. Avoid a "lemons market" by fostering trustworthy market dynamics

As explained in our response to the Inception Impact Assessment for Adapting liability rules to the digital age and circular economy, TFS considers that the long-term profitability of the AI industry in the EU will hinge on its ability to make the technology trustworthy and address the information asymmetry between providers and users of the AI system.[9] In brief, if the users cannot distinguish whether a system is trustworthy before buying it and if they do not feel protected enough from the potential risks of an untrustworthy AI system, their average willingness to pay for AI systems will decrease. Only the cheapest AI systems will therefore remain on the market in the medium term, due to consumers' low willingness to pay. If more trustworthiness requires additional R&D investment or design costs, this means that cheaper AI systems will on average be less trustworthy and therefore, over the long term, providers of trustworthy AI systems are "competed out" of the market (as their systems are too expensive). It results in a market with only untrustworthy AI systems that users do not want to pay for. This is an example of a "market for lemons" described by economist and Nobel Prize winner George A. Akerlof,[10] which is particularly applicable to AI systems. Further details are available in our response to the Inception Impact Assessment for Adapting liability rules to the digital age and circular economy.[11]

To offset this trend, we suggest the AI Act and broader governance system be adapted to ensure transparency and trustworthiness and therefore to keep the regulatory burden onto those with the most information available. As explained in-depth in our contribution on adapting liability rules to the digital age, increasing the responsibility of the actors with the most information available is the most efficient way for society to adopt the technology. The price mechanism will naturally reflect this increased responsibility and therefore ensures that incentives to innovate remain healthy.[12] The providers -which have the greatest effective control and full information on the design and development of the AI system- should therefore shoulder

[9] The Future Society - Liability rules for Trust in Excellence & Excellence in Trust , pp. 3-6
[10] The Market for "Lemons": Quality Uncertainty and the Market Mechanism
[11] The Future Society - Liability rules for Trust in Excellence & Excellence in Trust
[12] Ibid.

most of the responsibility, as is currently envisaged in the Act. Current discussions about the shift of obligations from providers to users are detrimental to a healthy and trustworthy market and would be harmful not only to consumers but also to industry - in particular domestic EU industry that is mostly composed of SMEs with limited lobbying and PR budgets (cf. our contribution for why these budget lines matter[13]). By attempting to reduce their regulatory costs, proponents of a shift of the obligations onto the users are reducing the overall size of the European market for AI systems and their future profits margins. Instead, correcting the information asymmetry between the provider and the user would be more beneficial - though much more costly, as explained in our previous contribution on Liability rules.[14] This is why preserving the obligations related to design and development (i.e. excl. obligations specific to the usage) onto the providers is strongly recommended.

Moreover, to maintain the highest level of trustworthiness, consumers should be empowered to lodge complaints directly with the relevant authorities, as per GDPR. In this sense, the possibility for parties with "a legitimate interest" to appeal conformity assessment decisions (Article 45) is welcome but limited to too few cases requiring extensive knowledge of the governance system. Likewise, while users are obliged to report any serious incident or malfunctioning, they must do so by contacting the provider itself rather than the national authorities (Article 29(4)). This slows down the flow of information relevant to preserving the health, safety and fundamental rights of EU citizens. To generate sufficient confidence in society, consumers need a more straightforward channel to the relevant authorities as well as the rights to not be subjected to non-compliant AI systems. Facilitating access to justice -notably for the citizens or small businesses least able to navigate the governance system- should be a priority for the protection of fundamental rights. This warrants a more direct complaint procedure rather than only the envisioned appeal and incident report procedures.

Finally, while we hope the complaint procedure will foster a climate of trust in the market, it remains an ex-post mechanism to repair violations rather than protect fundamental rights, health & safety. Ex-ante measures should therefore be included in the AI Act: for example, the recent controversy of Timnit Gebru and some of her colleagues' ousting[15] has highlighted the importance of providing a safe space for developers to share their concerns about AI systems their employer develops. This could be established by providing a direct link in the AI Act to the relevant EU measures "[protecting] persons who report breaches of Union law".[16] On the supplier-side, deployers and providers should be required to assess and disclose if an AI system is not aligned with the consumers or end users' interests. This "disloyalty" occurs when the system is designed to promote, sometimes temporarily, a potential third-party's interests above the user's interest, such as when a GPS reroute towards an advertising partners' restaurants rather than the user's preferred route or when the provider runs an A/B testing

[13] Ibid.
[14] Ibid. p. 3
[15] Wired - What Really Happened When Google Ousted Timnit Gebru
[16] Directive (EU) 2019/1937 of the European Parliament and of the Council of 23 October 2019 on the protection of persons who report breaches of Union law

experiment for a profiling company. The importance of loyalty in AI systems is crucial for consumer trust and there exist methods to assess this loyalty.[17] Moreover, there are contexts that increasingly leverage AI systems (therapy, financial advisory, …) where such conflicts of interest should be prohibited ex ante, as has historically been recognised in interactions and business transactions that do not involve AI systems.

Recommendations:

> Protect long-term market profitability from the information asymmetry's economic implications

- Preserve the AI Act's current allocation of obligations to the providers with regards to the requirements for which they have most information and effective control; prevent undue shift of obligations onto users;
- To the extent it can be achieved efficiently, facilitate the flow of technical information and understanding from the provider to its users.

> Empower citizens to defend their fundamental rights, health and safety more directly

- Require users to report serious incidents and malfunctioning to the relevant authorities, not only to the provider of the defective AI system;
- Establish a direct complaint procedure and extend individuals' right to not be subjected to non-compliant systems;
- Preserve the possibility for parties having a legitimate interest to appeal notified bodies' decisions.

> Provide ex-ante measures to foster more trust in the market

- Directly connect the AI Act to relevant measures in Directive (EU) 2019/1937 protecting whistleblowers;
- Require disclosure of the AI system's loyalty to users and factor in the importance of the loyalty of an AI system in the AI Act for consumers trust;
- Consider additional requirements for providers to foster trust as efficiently as possible.

Contact us at info@thefuturesociety.org for more information about this contribution.

[17] AI Loyalty: A New Paradigm for Aligning Stakeholder Interests