EU Proposed AI Legal Framework

Emre Kazim, Charles Kerrigan, Adriano Koshiyama
Holistic AI, CMS, UCL
Contact: emre.kazim@holisticai.com[1]

Abstract

The publication of the EU's draft AI legal framework is a milestone in the regulatory debate on AI. It proposes a risk based approach to regulating and reporting. In this white paper, we provide a high-level overview of the risk tiers, which we take to be the kernel of the legislation, and follow this by offering our initial thoughts and feedback on strategic points of contention in the legislation. Our main takeaways are: (i) Innovation - the sandbox approach may not be enough to ensure innovation; (ii) Reporting - in the lead up to codification we would like to see reporting being used to accelerate dissemination of best practice and benchmarking; (iii) Green-flagging - there does not appear to be sufficient detail to derive a reasonable set of green-flagging conditions; and, (iv) Manipulation - addressing the ambiguity in the draft proposal of banning systems with 'significant manipulation'. We conclude with notes on the legal status of algorithms, the status of GDPR in light of AI regulation and, the geopolitical ramifications of EU AI regulation.

Keywords

Regulation, Compliance, Artificial Intelligence, Accountability, Governance, Fairness, Transparency, European Union

1. Introduction

The publication of the EU's draft AI legal framework is a milestone in the regulatory debate regarding the rapid development and deployment of algorithmic systems by business and government in society. Such an intervention has been much anticipated, precipitated by the publication of a white paper on artificial intelligence by the EU in February 2020. The main takeaway from the resultant debates and consultation has been that a risk approach is most appropriate, with tiers that range from applications that are outright banned (c.f. Social scoring systems) to applications which require minimal transparency provisions (c.f. Spam filters). The

[1] **Emre Kazim** is Co-founder of Holistic AI, a start-up focused on AI-risk management. He is also a Research Fellow in Computer Science at UCL (emre.kazim@holisticai.com); **Charles Kerrigan** is a fintech partner at CMS (charles.kerrigan@cms-cmno.com); **Adriano Koshiyama** is Co-founder of Holistic AI and Research Fellow in Computer Science at UCL (adriano.koshiyama@holisticai.com).

draft legislation is lengthy and will result in sustained critique until the point of adoption and codification (it is envisioned that this will take two years).

In this white paper, we provide a high-level overview of the risk tiers, which we take to be the kernel of the legislation, and follow this by offering our initial thoughts and feedback on strategic points of contention in the legislation.[2]

Our main takeaways are:

- *Innovation:* the AI regulatory sandboxes may inadvertently present a significant barrier to innovation. We worry that it is likely to disincentivize research and development in areas that are likely to require sandbox access (by small scale enterprise) and because it is unclear what mechanisms will be in place, on the regulator's side, to (meaningfully) assess such speculative projects. Further, we believe universities should form a special category for exemption (one way to address this would be to stipulate that such risk management is for application rather than research). Our concern, otherwise, is that large companies will monopolies research and innovation.
- *Reporting:* the notion of reporting should be expanded such that for the period leading to codification of the legislation companies, researchers and other stakeholders should 'report' on issues surrounding documentation, problems faced, discursive statements around logic behind decisions etc.… to accelerate dissemination of best practice and, in relevant use cases, benchmarking
- *Green-flagging:* there does not appear to be sufficient detail to derive a reasonable set of green-flagging conditions. We interpret this as an opportunity for there to be a variety of risk-management frameworks, schemes and mechanisms, as well as reporting formats and procedures. Indeed, this is likely to become an important industry in itself.
- *Manipulation:* addressing the ambiguity in the draft proposal of banning systems with 'significant manipulation' is consequential to the ecosystem of innovation and trust the EU seeks to foster - in particular given the size and complexity of algorithmic systems in marketing.

2. Overview of Risk Tiers and Actions Therein

In this section we provide an overview of what we take to be the kernel of the draft legislation, namely the risk approach. Four tiers of risk are demarcated in the draft legislation. Below we begin with the lowest levels of risk (grouping limited and minimal risks) and end with the prohibited risk tier. For each risk tier we note the respective action provisioned.

[2] This contrasts to a thorough treatment of legal harmonisation i.e. integration with existing legislation, and to questions regarding implementation and enforcement across EU member states. Areas which we shall tangentially touch upon.

2.1 Limited and Minimal Risk

Limited risk concerns systems that do not pose a threat to the safety and livelihood of persons. Action in this context pertains to 'transparency obligations' i.e. users should be aware that they are interacting with a machine in order to make an informed decision (article 52). Here self-regulation and mechanisms of adhering to codes of practice are appropriate. *Minimal risk* concerns systems that do not pose a threat to the safety and livelihood of persons are considered. Here the right to opt out of the use of such technologies and transparency provisions (ex. ensuring users are aware they are interacting with a machine), suffice. No action is necessary in this context and it is envisioned that the vast majority of systems will fall into this category. We have grouped these together and will not treat them further as they are of least concern to our interests in this white paper.

2.2 High-risk

Here a general criterion is not offered, instead examples of sectors and applications are given (expanded upon below and corresponding to Title III, Chapter 2, articles 9-15). We infer from the case studies that, similar to unacceptable risk, such systems pose a threat to the safety and livelihood of persons, however, in these cases there are benefits that can be derived and used to justify deployment through good governance/risk management.

- *Action:* Risk-management system and reporting

In Annex III to the draft legislation, a list of high-risk systems, along with explanandum, is given. In summarised form, we reproduce this below.

- *Biometric identification and categorisation of natural persons:*
    - 'real-time' and 'post' remote biometric identification.
- *Management and operation of critical infrastructure*
    - safety components in the management and operation of road traffic and the supply of water, gas, heating and electricity.
- *Education and vocational training*
    - (a) determining access/assigning to educational/vocational training institutions;
    - (b) assessing students in educational and vocational training and for assessing participants in tests commonly required for admission to educational institutions.
- *Employment, workers management and access to self-employment*
    - (a) recruitment/selection of persons, notably for advertising vacancies, screening or filtering applications, evaluating candidates in the course of interviews or tests;
    - (b) making decisions on promotion and termination of work-related contractual relationships, for task allocation and for monitoring and evaluating performance and behavior of persons in such relationships.
- *Access to and enjoyment of essential private services and public services and benefits*
    - (a) public authorities or on behalf of public authorities to evaluate eligibility for public assistance benefits and services, as well as to grant, reduce, revoke, or reclaim such benefits and services;
    - (b) evaluate creditworthiness or establish credit score, with the exception of AI systems put into service by small scale providers for their own use;

    - (c) dispatch, or to establish priority in the dispatching of emergency first response services, including by firefighters and medical aid.
- *Law enforcement*
    - (a) making individual risk assessments in order to assess the risk for offending or reoffending or the risk for potential victims of criminal offences;
    - (b) polygraphs and similar tools or to detect emotional state;
    - (c) detect deep fakes;
    - (d) evaluation of the reliability of evidence in the course of investigation or prosecution of criminal offences;
    - (e) predicting the occurrence or reoccurrence of an actual or potential criminal offence based on profiling or assessing personality traits and characteristics or past criminal behaviour of persons or groups;
    - (f) profiling for detection, investigation or prosecution of criminal offences;
    - (g) crime analytics regarding persons, allowing law enforcement authorities to search complex related and unrelated large data sets available in different data sources or in different data formats in order to identify unknown patterns or discover hidden relationships in the data.
- *Migration, asylum and border control management*
    - (a) polygraphs and similar tools or to detect the emotional state;
    - (b) assess a risk, including a security risk, a risk of irregular immigration, or a health risk, posed by a person who intends to enter or has entered;
    - (c)verification of the authenticity of travel and supporting documents and detect non-authentic documents by checking their security features;
    - (d) examination of applications for asylum, visa and residence permits and associated complaints with regard to eligibility of persons applying for a status.
- *Administration of justice and democratic processes*
    - researching/interpreting facts and law, in applying the law to a set of facts.

Although no definition is given, it is clear that ***high-risk overlaps concerns with incursions to civil rights, participation, access and due process***. In each case the corresponding article is highly abstract and we read articles as indicating what needs to be of concern rather than narrowly and prescriptively outlining risk mitigation.

In such high-risk cases a number of (legal) requirements are stipulated in terms of justifying the use of these high-risk systems. Indeed article 9 asserts the need to establish a 'risk management system' that must be acted upon and maintained, including adequate documentation. It is suggested that this is a 'continuous iterative process run throughout the entire [high-risk system's] lifecycle'. Following this, articles 10-15 denote, in more detail, the conditions that have to be met for a system to be justified for use.

- *Data and data governance (article 10):* Training, validation and testing data sets to ensure that they are of high quality data.

- *Documentation (article 11, 12):* Provide detailed documentation for third party assessment, including technical documentation and record-keeping i.e. period logging of standards specifications being met.
- *Transparency for users (article 13)*: comprehensible information regarding contact details of provider, purpose, accuracy, security, data used, human-oversight measures and expected life-cycle of a system should be reported.
- *Human-oversight (article 14):* Must ensure high-level of human oversight in development and deployment, through appropriate interfaces. The overseers must be able to understand the capacities and limitations of a system, avoid automatically accepting recommendations of a system, and be able to intervene effectively. Decisions should be taken after at least two people have overseen the system.
- *Accuracy, robustness and cybersecurity (article 15):* such relevant metrics must be declared, including failsafe mechanism, mitigation strategies against vulnerabilities and for cybersecurity attacks.

2.3 Unacceptable risk

Here concern is with systems that pose a direct and clear threat to the safety, livelihoods and rights of people.

- *Action:* outright ban.

Three use cases are named, these are:

- *Social Scoring systems:* in opposition to systems that have been used in China, inferring character judgements from social behaviour is banned. Cases where a person incurs traffic incidents or engages in other kinds of antisocial behaviour should have no bearing on other (public) services/benefits they may receive.
- *Manipulation:* two kinds of algorithmic manipulation are discussed, namely that of vulnerability and that of the subliminal. Regarding the former, the elderly, children, and those with disabilities are noted. Regarding the former, we read reference to the ‘algorithmic nudging literature, which concerns a person being ‘nudged’ i.e. manipulated by an algorithm towards a particular end, such as voting for a political party/candidate or purchasing. A qualifier ‘significant’ is introduced, however, it is noteworthy that no sustained treatment of what constitutes significant/subliminal manipulation is given.
- *Remote biometrics:* the use of indiscriminate scanning and use of identifiable characteristics (c.f. Facial recognition, audio scanning, sentiment analysis in the public sphere, etc.) are banned. The qualifier ‘remote’ is used to indicate that individual and consensual use of such systems is fine i.e. logging in via fingerprint, face, voice, etc.

It is noteworthy that exceptions are made with respect to special cases of law enforcement. Examples are given such as averting a major security threat and searching for missing persons.

3. Comments

In this section we offer our comments based on initial reading of the proposed legislation. Note that we welcome the proposal as a first, and much needed step, in the standards and regulation of AI, and that our commentary will focus more on areas of concern rather than areas that we endorse.

3.1 Innovation

AI is being rapidly adopted across business and government because it offers immense potential in terms of how services and products are delivered, and with respect to the kinds of services and products that emerge as a result of the technology. In this respect, AI is a tremendous economic opportunity that is driven by research and innovation (something that is often referred to in the proposed legislation). As such, it is critical that innovation is at the forefront of any AI governing framework - or rather the need to ensure that innovation is not stifled. In the proposed legislation articles 53-55 address the issue of innovation under Title V 'Measures in support of innovation'.

Here the main takeaways are that *AI regulatory sandboxes* (article 53) should be established by member states as environments that facilitate the 'development, testing and validation of innovative AI systems for a limited time before their placement on the market or putting into service pursuant to a specific plan'. Processing of personal data (article 54), or other sensitive data for public interest applications is mentioned specifically. Title V concludes with article 55 'Measures for small-scale providers and users', where it is asserted that 'small-scale providers and start-ups [should be provided] with priority access to the AI regulatory sandboxes'.

There are two concerns with this:

- *Sandboxes as additional burden:* notwithstanding the assertion regarding priority access, in practicality what will in fact be the effect of needing to conduct research and development in such sandboxes for start-ups and small scale providers? We envision that this will present a significant barrier to innovation because it is likely to disincentivize research and development in areas that are likely to require sandbox access and because it is unclear what mechanisms will be in place on the regulator's side to meaningful access such speculative projects.
- *Universities:* we had anticipated an exception for university research and according to our reading no expectation is discussed. This is striking as much of the cutting edge research occurs in this context. Furthermore, the universities are key players in stimulating the entire AI innovation and economy, and thus exception/special provision is to be expected. One way to address this would be to stipulate that such risk management is for application rather than research.

We worry that the consequence of these two concerns will mean that large companies will be the only ones with sufficient resources to satisfy regulatory burdens, further monopolizing the

technology market and facilitating rapid movement to such companies of engineers and scientists from the academy who want to undertake the most innovative research.

3.2 Reporting

Title VIII (articles 61, 62, 66) sets out the monitoring and reporting obligations in a number of contexts. Most notable is article 62 which outlines reporting in contexts of serious incidents and of malfunctioning. We believe that the notion of reporting should be expanded such that for the period leading to codification of the legislation companies, researchers and other stakeholders should 'report' on issues surrounding documentation, problems faced, discursive statements around logic behind decisions etc… We argue for this because we believe that practical experience 'from the field' is most valuable and it is an imperative that knowledge transfer is facilitated. Should a mechanism be established (perhaps in the form of an anonymous forum or via facilitated workshops dedicated to such knowledge transfer or as a repository) this would accelerate dissemination of *best practice* and, in relevant use cases, *benchmarking*.

3.3 Green-flagging

As noted above, no formal definition of high-risk is offered, no set of necessary conditions are outlined etc. instead sectors and case studies are listed and described. Notwithstanding this it is clear that impact upon civil liberties, due process and discrimination can be drawn out as themes, and with that a coherent definition/set of conditions can be derived from these case studies. In contrast to this, although mitigation is discussed in abstract terms (risk governance, data governance, reporting etc.), there does not appear to be the same level of commentary needed to derive a reasonable set of green-flagging conditions. *We interpret this as an opportunity for there to be a variety of risk-management frameworks, schemes and mechanisms, as well as reporting formats and procedures. Indeed this is likely to become an important industry in itself.*

3.4 Manipulation

Systems used to significantly manipulate people are banned. However, with terms such as 'subliminal' in the text and with little qualification of what constitutes significant manipulation, this is an area that is likely to be read ambiguously. That a system targets, curates, and recommends for an individual and/or group is the case in all forms of marketing (and other service provision) which has the end of realising a purchase. Where the line is drawn, with respect to what this constitutes, is contentious, as well as what falls within the domain of systems of manipulation (ex. How a person is driven to navigate through to purchase, to 'click' or 'link', what options are presented, if image or text or audio (or a combination thereof) is used, if protected characteristics are used for an individual or through inference and aggregation, etc…). Given the size and complexity of algorithmic systems in marketing, we believe addressing this is consequential to the ecosystem of innovation and trust the EU seeks to foster.

4. Conclusion

Welcoming the proposed EU AI legislation as a concrete step in the codification of standards, we anticipate significant revision around the areas we have noted in section 3. Other areas that we anticipate will be points of debate also raise some legal issues. These include:

- *The legal status of algorithms:* analogous to the way that companies are granted 'personhood' - will it be necessary to grant personhood to certain AI? Incorporation was an innovation centuries ago that enabled a non-natural person to acquire legal personality so that it could hold assets, enter into contracts and sue and be sued. The most common form of incorporated entity is a company. Importantly, shareholders in a company do not bear responsibility for the actions of the company because they are separate entities (a principle commonly known as the "corporate veil"). Over time, rules relating to incorporation have changed to include more entities (for example incorporated partnerships in addition to private companies limited by shares) and to provide some exceptions to the corporate veil principle. Most recently, DAOs (decentralised autonomous organisations) using smart contracts have extended the analysis of the nature of entities with legal personality in ways that mean that there is now scope to extend this further to algorithms. Alongside this, questions of responsibility for the acts and omissions of an algorithm will need to be considered. While it is conceivable that policymakers may permit algorithms to have legal personality this will only be the case initially if a person responsible for financial and non-financial losses is identified in every case.
- *The status of GDPR in light of AI regulation:* although there is an imperative to harmonise with GDPR, as we argue elsewhere (Kazim et al, 2020), it may be necessary to revise GDPR in light of AI. Again, recent developments in other types of emerging technologies have tested the rules written for GDPR. The conflict between the "right to be forgotten" and the immutability inherent in permissionless blockchain technologies is well known. Jurisdictions that run developments in data and AI regulation together are most likely to find the best models to apply. GDPR is still only a limited success in terms of compliance and in terms of user experience for the consumers that are intended to benefit from it. In fact, this question and the one above may be addressed together. As personal devices are embedded with more sophisticated AI, users' own AI algorithms will likely take on the role of managing data privacy and interactions with other entities that the user encounters both in a commercial and in a personal context.
- *Geopolitical ramifications of EU AI regulation*: there is a sense in which the EU aims to be 'the' AI regulation – similar to how GDPR became de facto 'the' privacy provision. However, we believe that there is an asymmetry here: in terms of GDPR there was a framework of protection (so it was easy for others to adopt) but AI is a product and resource itself (so much less likely to be adopted by others) as this would be akin to the sharing of national resources. The EU approach is generally consumer-centric. It is likely that other jurisdictions will take different approaches. The most important point over time will be to what extent the outcomes benefit users and how the trade offs in each case operate.

Selected Bibliography

- Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. European Commission. Brussels, 21.4.2021. COM(2021) 206 final. 2021/0106 (COD) https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence
- Annexes to the Proposal. European Commission. Brussels, 21.4.2021.COM(2021) 206 final. ANNEXES 1 to 9. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence
- Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., ... & Lomas, E. (2021). Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3778998
- Kazim, E., & Koshiyama, A. (2020). The interrelation between data and AI ethics in the context of impact assessments. AI and Ethics, 1-7. https://link.springer.com/article/10.1007/s43681-020-00029-w