Feedback on the EU AI White Paper

Author: Ron Bodkin

please note that this commentary represents my personal opinion and in no way represents the opinion of my employer

Date: May 26, 2020

I wrote this essay to share my perspective on the European Commission's White Paper On Artificial Intelligence - A European approach to excellence and trust from February 2020. It's laudable that the EU is defining policy to get ahead of the challenges of misuse of AI, while still seeking to foster the development and beneficial uses of AI.

Overall, I would like to see more of a focus in the white paper on addressing immediate problems that have been observed in AI systems with regulation to change behavior. I also believe there's a need for more flexible means of regulation and audit to adapt and respond to problems stemming from AI. My concern is that the white paper punts on concrete regulation of most uses of AI and indeed defines a narrow scope of high risk areas.

AI Risks

In particular, consider the scope of risks contemplated:

> *The main risks related to the use of AI concern the application of rules designed to protect fundamental rights (including personal data and privacy protection and non-discrimination), as well as safety and liability-related issues.*[1]

I believe this is too narrow and that risks and harms should be based on impact of algorithms and evidence of current harm. In particular, SMEs are likely to have significantly less impact and therefore only more severe individual risks or harms should be regulated for them (e.g., those involving human safety). The impact of AI should also be assessed by way of comparison of processes before AI and after AI was introduced to determine incremental impact rather than comparing AI systems with perfection.

Moreover the white paper defines operational criteria for high-risk AI applications as meeting "the following two cumulative criteria:"

> *First, the AI application is employed in a sector where, given the characteristics of the activities typically undertaken, significant risks can be expected to occur. This first criterion ensures that the regulatory intervention is targeted on the areas where, generally speaking, risks are deemed most likely to occur. The sectors covered should be specifically and exhaustively listed in the new regulatory framework. For instance, healthcare; transport; energy and parts of the public sector. The list should be*

[1] Page 10

periodically reviewed and amended where necessary in function of relevant developments in practice;
Second, the AI application in the sector in question is, in addition, used in such a manner that significant risks are likely to arise. This second criterion reflects the acknowledgment that not every use of AI in the selected sectors necessarily involves significant risks. For example, whilst healthcare generally may well be a relevant sector, a flaw in the appointment scheduling system in a hospital will normally not pose risks of such significance as to justify legislative intervention. The assessment of the level of risk of a given use could be based on the impact on the affected parties. For instance, uses of AI applications that produce legal or similarly significant effects for the rights of an individual or a company; that pose risk of injury, death or significant material or immaterial damage; that produce effects that cannot reasonably be avoided by individuals or legal entities.[2]

In particular, I will review a number of cases that should be addressed urgently. I see these examples as violating fundamental rights but not in the areas emphasized in the definition (personal data, privacy and discrimination). Moreover, these problems arise in technology, media and commerce which are not examples given as high risk areas for AI applications. A root cause of many of these harms is unintended consequences of optimizing systems to maximize revenue - whether to sell ads or subscriptions through addictive engagement or to manipulate users or to abuse privacy for commercial gain. Indeed, the EDPS Opinion on online manipulation and personal data identifies numerous areas of deception and manipulation caused by AI. I believe these must be addressed by government regulation, not hope that industry self-regulation will cure these problems.

In general, AI algorithms are having a significant impact on the media, political, social and psychological environment. They are also increasingly central in shaping and manipulating consumer behavior. Consider research like Exploring Echo-Systems: How Algorithms Shape Immersive Media Environments and the references contained within it and this analysis from a former YouTube recommendation engine engineer on how AI algorithms are driven by a focus on short-term engagement and are implicated in many of the harms listed above.

In particular these are areas I view as high risk that should be explicitly addressed (drawn from my Position Paper on Responsible AI):

- AI algorithms' role in election manipulation, e.g., targeted voter suppression by Cambridge Analytica and AI algorithms that were manipulated by Russia to affect elections. A root cause of these problems is a lack of inadequate controls and security in AI systems as well as a conflicting incentive whereby AI algorithms promote this kind of inflammatory content because it increases engagement, hence advertising revenue.
- AI algorithms driving digital addiction - a combination of AI algorithms and design techniques directly or indirectly maximize time users spend to their detriment. These

[2] Page 17

algorithms are prevalent in social media, video and other media sites and video games. In general digital media properties optimize for user engagement, e.g., YouTube announced in 2012 that it had shifted from optimizing from click through rates to watch time and FaceBook announced in 2018 that it was prioritizing "time well spent." TikTok has recently become a major influence on youth with an experience driven by AI recommendations. TikTok has said little about their recommendation system, but researchers have observed racial, political and LGBT bias in these algorithms. Moreover, there is a consistent strong commercial motivation for advertising or subscription business models to drive engagement through AI algorithms.
The consequences are significant. Increased screen time is implicated in teenage depression and suicide as in this study. Increased time spent on social media has been shown to be associated with increased feelings of social isolation. A recent survey has shown how prevalent feelings of regret by users are about the apps they use - and that regret is highly correlated with the time users spend. This is a strong indication that users are being manipulated into addictive behavior against their intentions by prevalent AI algorithms that are increasingly dominant as a form of leisure activity.

- Consumer deception and manipulation. Consider Amazon manipulating their AI algorithm-driven search results to maximize their profitability while violating user expectations that organic search results are based on their preferences, as reported by the Wall Street Journal. I see this as a deceptive practice and a kind of unacceptable manipulation. While promoted items are a well understood concept, using AI to bias organic search results should at a minimum require transparency (just as distinguishing ads from content is important in media). This is especially impactful when done by the leading e-commerce seller in most Western EU countries (e.g., see this eMarketer analysis) and one that claims its top value of "customer obsession" means it seeks to "earn and keep customer trust."
- Misinformation and extremist content - AI recommendations have been observed to promote extreme or radical content like conspiracy theories, hate speech, misinformation and Holocaust denials as reported by Zeynap Tufecki about YouTube. AI-driven content recommendations have led to problems like vaccine hesitancy and leading to an erosion of trust in facts and science that is the foundation of democratic society. The recent changes made by FaceBook and YouTube to address these issues is indicative of the severity of impact from these problems. In extreme cases, these problems have led to clear human rights violations such as FaceBook's news feed AI algorithms promoting inflammatory content that led to mass killings in Myanmar. But it shouldn't require that kind of extreme outcome for us to be concerned about impacts of AI on a free society.
- Filter bubbles driven by AI algorithms are a related societal challenge. This reduces common ground that's foundational to a functioning democracy. Arguably, this has led to erosion of public trust in science in response to the Coronavirus pandemic and resulted in reckless behavior that spreads a lethal disease and prolongs the economic damage.
- AI to manipulate consumers to buy more expensive items. E.g., this article about credit cards using purchase history to sell. Should consumers have recourse to limit how recommendations influence their decisions? Should this require opt-in?

Enforcement

It would be valuable to explicitly address and require mandatory regulation to address the kinds of harm stemming from AI usage in technology, media and commercial sectors discussed above. Yet the white paper and the previous EDPS opinion imply that voluntary guidelines for self-regulation will suffice. I believe that there needs to be a more flexible, dynamic regulation model.

The white paper calls for “Initiatives could also include the support of sectoral regulators to enhance their AI skills in order to effectively and efficiently implement relevant rules.”[3] I would recommend constituting a similar construct to national Data Protection Authorities that are responsible for governing trustworthy AI within European states consistent with an overall rule. I also think that it’s important to improve the knowledge and responsiveness of regulators by requiring a private AI audit function to certify compliance for moderate or high risk applications of AI rather than requiring complaints and investigations by a regulator as the primary means of control. This would allow auditors to compete for efficiency while proving effectiveness of audits. This is akin to Clark and Hadfield’s proposed approach of regulatory markets for AI safety. At a minimum, having more scope for private innovation to advance rules would be of great value given the speed of change in this sector and challenges in recruiting top talent for regulatory purposes - this is important both to limit harms from AI misuse and to ensure clarity for what is allowed to advance the benefits of AI.

The white paper further says “Given how fast AI is evolving, the regulatory framework must leave room to cater for further developments. Any changes should be limited to clearly identified problems for which feasible solutions exist.”[4] An auditor function can reduce the challenges with opacity and can allow for auditors to certify best practices based on state of the art capabilities rather than requiring frequent updates to dynamic regulations. Requiring notification of discovered failures to comply in confidence to regulators (like security failures) would be valuable in any event to track known problems - this can be decoupled from consequences where failures arose based on good faith efforts applying reasonable techniques.

The white paper notes “Enforcement authorities and affected persons might lack the means to verify how a given decision made with the involvement of AI was taken and, therefore, whether the relevant rules were respected.”[5] - this is a reason why an audit function for high risk or high stakes algorithms should be required where auditors must meet standards for expertise and be liable for failures. Also, by allowing companies to engage auditors they can find those who work best with them and their approaches.

---

[3] Page 6

[4] Page 10

[5] Page 12

Consider further this aspect of enforcement for safety-critical AI systems

> *A lack of clear safety provisions tackling these risks may, in addition to risks for the individuals concerned, create legal uncertainty for businesses that are marketing their products involving AI in the EU. Market surveillance and enforcement authorities may find themselves in a situation where they are unclear as to whether they can intervene, because they may not be empowered to act and/or don't have the appropriate technical capabilities for inspecting systems*[6]

Having an auditor certify safety would reduce risks and uncertainty for those creating AI technologies. The auditor would be responsible for demonstrating safety in a liability lawsuit and auditors might be required to carry minimum liability insurance.

The white paper does not contemplate mandatory requirements for transparency nor for accountability for leaders in companies that deploy high risk AI. Having a certification approach that allows for relatively low cost updates would be valuable (e.g., with an audit function).

In a deep way, the increasing concentration of power by large technology companies with narrow commercial aims and a lack of access or ability to access by the public is problematic. The use of AI editorial allows for increased secrecy and increases the impact and speed of harms whereas a human-driven system is subject to more judgment and to whistle blowers identifying harms to prevent damage. In these areas, while using AI to block extreme content is relevant and has benefits and harms, most of the impact of AI comes from what content is amplified and promoted. I believe that legislation to allow for competition for algorithms and diversity of perspective is highly desirable, even as technology platforms increasingly limit the ability of third parties to offer alternative experiences that interact with their proprietary networks, data, and systems.

Other Considerations

In addition to training AI developers (as described on page 6), it would be valuable for EU policy to train other stakeholders on AI including how it works, oversight and accountability in a system, e.g., for business leaders, product management, user experience professionals, social scientists, and risk and compliance professionals

Pages 8-9 describe a fairly narrow scope of interest for how the EU can respond to the threat of harmful AI development and deployment of systems by companies subject to authoritarian or other governments with policies that obligate contravention of these policies. E.g., consider government surveillance of EU citizen activity or requirements for companies to comply with privacy violations for EU citizens. The white paper does not propose means to rectify these harms beyond the narrow scope of unequal terms of trade in data access.

[6] Page 13

Conclusion

In conclusion, I think it’s important to define a flexible but binding regulatory framework that can address the dynamic, fast moving nature of AI. It’s important to understand how AI systems are changing our society including our democratic institutions and media and how they are affecting the public. With the right scope, the EU’s rules for AI can be a powerful force to course correct unintended harms while allowing great benefits to society.

I believe there needs to be a balance between risk and cost of regulation, also a recognition that it's hard to assess the impact of a technology in advance, so transparency in monitoring and responding to problems will be important.