Fonden
TEKNOLOGIRÅDET
DANISH BOARD OF TECHNOLOGY

Bringing voices from society to bear on the EC white paper On Artificial Intelligence - A European approach to excellence and trust

The feedback letter is written by senior project manager Lise Bitsch and project manager Nicklas Bang Bådum.

With the publication of its white paper on artificial intelligence[1], the European Commission also opened for a public consultation. The Danish Board of Technology Foundation (DBT) would like to thank the EC for the opportunity to provide our input.

The commission white paper sets itself apart by highlighting trustworthiness as a central value, and design principle in the European strategy. If we are to use AI in areas such as transportation, courts of law and systems of justice, police, or healthcare systems, then strict rules must apply for the control with and security of such applications. AI has a huge potential for helping bring about improvements and positive change in our societies. That is why the potential pitfalls are also very far ranging. The white paper integrates recommendations and guidelines from the EU AI high-level group (AI HLEG), and it clearly condemns uses of AI that are discriminatory or in other ways violate human rights. The white paper highlights the need for the development of AI to continue in a controlled manner and with respect for the humans using and affected by the implementation of AI-based solutions.

Our feedback is based on our experience working with societal, legal, and ethical aspects of AI in the context of several European projects[2], and our more than 30-years of experience working with citizen engagement in policy and innovation processes. On that basis we make the following recommendations for consideration in the completion of the EC white paper on AI[3]:

[1] European Commission, White Paper on Artificial Intelligence: a European approach to excellence and trust, 19 February 2020. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust

[2] The projects are: The EU flagship 'The Human Brain Project' and EU projects PANELFIT and SurPRISE

[3] Central to our reply are findings from two recent activities undertaken in the context of the EU flagship Human Brain Project (HBP).

- ***"AI 360 Copenhagen",*** was a workshop held on March 21-22, 2019. In the workshop an international group of experts in ethics, law, social science, culture, politics, and economy for a multi-dimensional and thorough treatment of AI and its implications for our future societies. Over the course of two days they deliberated and prioritised the most concerning possible implications of AI across the dimensions "Politics", "Rights and Ethics", "Law and legal frameworks", "Economy", and "Society", and they developed recommendations for how to solve the concerns they had themselves pointed out.
- ***"EuropeSay on AI"***, were Europe-wide citizen dialogues, engaging more than 900 citizens in 13 countries in deliberation on uses of AI. The citizens took the outcomes of the AI 360 Copenhagen workshop as their starting point to point to the areas they found the most critical to the future development of AI, and to develop their recommendations on how the development of AI should be governed.

1. Undertake work to uncover and address the full range of potentially abusive uses of AI, including societal, political, security, intelligence, and military domains
2. Make obligatory regulation for all AI
3. Reconsider the distinction between high and low risk AI
4. Encourage public authorities to carry out a comprehensive investigation of their local challenges before implementing AI solutions
5. Build trust by supplementing balanced information with involving citizens in deliberation, agenda-setting, prioritization and decision making
6. Encourage national and European deliberation and action on the role of digital spaces of interaction, news and debate central to the functioning of our democracies, human rights and well-being of European citizens
7. Human oversight should be complimented by human insight and explainability

Undertake work to uncover and address the full range of potentially abusive uses of AI, including societal, political, security, intelligence, and military domains

If developments in AI live up to the expectations expressed in the EU white paper, it is hard to imagine any part of our lives that will remain untouched. The emergence of AI opens perspectives of change and improvement in healthcare, in dealing with climate change, in building more effective farming systems and providing increased security for Europe and its citizens. As the white paper rightly notes, such developments also open questions on the abuse of AI for criminal purposes. However, it goes on to note that issues of misuse for criminal purposes and any provisions to tackle such issues are outside of the remit of the white paper. In addition, the white paper does not make any mention of addressing other potential misuses or abusive uses of AI. But this is an essential aspect that needs to be addressed.

Taking the example of abuse in the domain of politics, AI could be used to support abusive and manipulative practices in e.g. democratic election processes. AI-enforced microtargeting techniques could make it impossible for voters to judge the coherence of policies, positions of political parties, politicians, and arguments. Also, techniques for fostering division between societal groups, e.g. by pushing specific messages about one social group to another, could become more widespread.

Technology is used in social contexts, and can be part in enhancing our emotions, convictions, and beliefs. Part of the challenge in developing trustworthy AI, is that it needs to function in social and societal contexts. Today, extremists use e.g. social media and digital platforms to create and maintain communities where discriminatory, distorted, and abusive messages and material is shared. We have also seen how online forums helped suicidal individuals nurture suicidal thoughts and even help each other commit suicide. Additional questions include use of AI by business to influence consumer behaviour, or use of AI solutions by insurance companies, and the limits of such uses to avoid increased inequalities in e.g. health.

In the military and intelligence domains, AI could become a weapon in for terrorism and cyberwarfare. The infrastructure of nation states would need strong protections against cyber-attacks. Internationally, open,

and trustworthy channels of communication between states could be essential to maintain trust and counter effects of misinformation.

The white paper does not mention concrete plans for taking up issues of misuse, and therefore we mention the urgency to do so here. Issues of the possible abusive uses of AI, does not only concern AI used for criminal purposes, but also the potentially abusive applications of AI in the societal, political, security, intelligence, and military domains. The potential for abuse is one of the most concerning aspects about the future application of AI[4]. In developing future AI systems and applications there is therefore a need to develop an overview of the full range of potentially abusive uses of AI, and to develop strategies for handling them. Better understanding and tackling these challenges are essential to the future functioning of our democratic systems.

Reconsider the distinction between high and low risk AI

The ubiquity of potential areas of application is testament to the potential of AI. And while it might seem sensible at first to differentiate between AI systems and services based on the risk that they comprise, and to use this distinction to decide which systems to regulate, such distinctions are almost always only meaningful in a limited way.

First, because almost no application of AI is without risks, many of which can be severe. Take the example given in the white paper text of a flaw in a hospital appointment system[5]. At face value AI supported booking systems in hospitals, and other sectors, seem unproblematic. However, in hospitals, patients need to be able to trust that the information provided the system is processed, transmitted, and stored securely, and that it is only accessed or transmitted to the right people. The severity of a data breach or other security breaches in such a system, is already a reason why it should be subject to legislation. In addition, it is not difficult to imagine the integration of such a system with an AI based diagnostics support tool. One could imagine such a tool would suggest diagnoses based on the input given by the prospective patient, but also help prioritise the order of consultations. The combination of data points and variable uses of data exacerbates the need to ensure the security and efficacy of the system.

Most AI systems will not operate in isolation: More likely they will become integrated in a wider ecosystem of multiple systems interacting. For this reason, what is a low-risk system in and of itself, can in connection with other systems provide one piece in a system which is no longer low-risk.

Additional questions include, who it is that make the assessment of the risk factor of an AI system or service, and how the requirement of having the high-risk AIs undergo compliance checks is to be enforced.

[4] Chapter 5 and 6 in report from the AI 360 workshop : https://sos-ch-dk-2.exo.io/public-website-production/filer_public/25/81/2581b77f-c39b-4359-b835-4f5163214378/ai360_humanbrainproject_recommendations_report_final.pdf

[5] European Commission, White Paper on Artificial Intelligence: a European approach to excellence and trust, 19 February 2020. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust , page 17

There are at least two options for an assessment regime based on the high and low risk AI distinction. Each has its drawbacks. The first is that the companies, developers etc. that develop or market the AI system are responsible for making the assessment and submitting the AI application for review. It has the advantage that not every system or service has to be checked. It also gives companies working on 'low risk' systems lower barriers to entry into the market. However, if the companies are to carry out self-assessments, the odds are that the number of high-risk applications might be (unrealistically) low. The second option is that an independent public authority makes the assessment, which will be a more effective and reliable approach, but at the same time will also be more burdensome and resource demanding for companies. Furthermore, there is the question as to the frequency with which the assessment is to be made. Should it be a one-off assessment, or should an application be assessed multiple times, e.g. when uses are expanded or the system is updated?

For this reason, we recommend discarding the high-low risk distinction, and when assessing the risks associated with AI systems it is for the above mentioned reasons also necessary to consider the risk of function creep and repurposing of the system.

Make obligatory regulation for all AI

The white paper identifies the potential difficulty of enforcing some fundamental rights and applicable legislation on AI, because of the technology's intrinsic features, e.g. the tendency for AI analyses to be opaque[6], its complexity, unpredictability, and partially autonomous behaviour[7]. In addition to this comes issues of allocation of responsibility and liability. Further, the white paper cites the High-Level Expert Group's feedback process, which found that some of the requirements suggested in the HLEG's The Ethics Guidelines for Trustworthy Artificial Intelligence, are not specifically covered by existing legislation, namely transparency, traceability and human oversight.

The white paper suggests that only AI considered high-risk should be subject to AI specific mandatory regulation. However, EU Citizens are almost universally in agreement that all AI should be regulated[8]. The most popular suggestion was a mandatory case-by-case approval of any AI application, second to which was that a public, government or EU authority should issue a certification of good practice that companies can apply for. These two were close to being equally popular, so there is no clear-cut recommendation as to what the regulation should look like. However, some experts agree, that AI should be subject to mandatory regulation[9] , and that this regulation could very well be the case-by-case regulation chose most frequently by citizens[10].

Therefore, we recommend to implement a system of mandatory regulation for all applications of AI.

[6] Ibid, pp. 10 and 12
[7] Ibid. p. 12
[8] EuropeSay on AI results report
[9] ALLAI: *First Analysis of the EU White Paper On AI* (https://allai.nl/first-analysis-of-the-eu-whitepaper-on-ai/)
[10] AI360 results report

Encourage public authorities to carry out a comprehensive investigation of their local challenges before implementing AI solutions

AI is already revolutionising many workflows and established professions. AI is also imagined bringing about many improvements in many areas of the public sector such as social services, healthcare, infrastructures, budget spending and allocation. However, before deciding on a general application of AI in the public sector, it is necessary to carefully consider the specificities of that sector and the social, cultural, and political context that the AI application is to function in. In the public sector, decisions often affect individual persons or groups of persons, and careful analysis of the problem at hand, and any possible discriminatory or otherwise undesired side effect should be carefully considered before deciding on implementation of AI. AI should thus not be treated as a silver bullet. While it does indeed hold widespread potential and innovation in the products will be propelled further by the spread in its uptake, it is important to consider alternative solutions as well.

Instead of starting with the solution (the AI system), it is crucial to start with considering what the problem is, and what the possible solutions could be, including but also beyond AI. Sometimes the AI system will indeed be the best solution, whether this is determined based on cost, speed, efficiency, or other criteria.

Building trust by supplementing balanced information with involving citizens in deliberation, agenda-setting, prioritisation, and decision making

The two main building blocks of the EU white paper are the building an ecosystem of excellence and an ecosystem of trust for AI. The adaptation and consequently innovation in AI systems, products and services depend on a societal acceptance and trust in the systems. However, the white paper does not address how to achieve a public sentiment of trust in AI.

In our EU-wide dialogues with citizens, we found that in general AI is considered favourably. The more knowledge on AI citizens professed beforehand, the more likely they were to also endorse AI[11]. However, there are still considerable populations of citizens reporting they have yet to form an opinion about AI. Therefore, information and education of the wider population about AI could be important means of creating and maintaining public support for research, development, and application of AI. Furthermore, moving forward it is important to provide the general population with reliable and balanced information on AI and its uses.

In addition to communication and information on AI, citizens should be heard in the processes of deciding on contexts of use of AI. The white paper states that it intends to address the concerns of citizens. To this end, inviting NGOs and consumer organizations is a useful means and it will help broaden the understanding of the issues at stake. However, these organizations cannot represent the full range of opinions and values of EU citizens. To truly build an ecosystem of trust, citizens must have a place at the table and influence on decisions made in relation to AI regulation and use. The white paper lists several assumptions on what citizens are concerned about, but it provides little background on where these concerns are derived from.

[11] Results report from EuropeSay on AI citizen engagement

The concerns mentioned in the white paper are indeed important aspects. Additional concerns, that we found in our dialogues with citizens, include concerns on unintended effects of AI (support)systems, use of AI for malicious purposes, fears of being left powerless in defending own rights and safety, and concerns over facing information asymmetries of algorithmic decision-making[12].

Increased engagement with citizens will provide more information and insight into the needs, concerns and values that are important to them in the development of trustworthy AI in Europe. Such dialogue can fruitfully also be designed to encompass agenda setting for publicly and EU funded research, as well as discussions about what uses are acceptable or unacceptable, what uses are controversial and under which conditions these could be acceptable. This will not only create trust in the systems and the EU, it will also create buy-in among consumers because they will feel that they have been listened to and their concerns have been taken into consideration. Citizens are both willing and able to engage in nuanced discussions about complex topics, and that such processes can produce high-quality output for e.g. new research agendas[13].

We therefore recommend building trust by supplementing balanced information with involving citizens in deliberation, agenda-setting, prioritisation, and decision making.

Encourage national and European deliberation and action on the role of digital spaces of interaction, news and debate central to the functioning of our democracies, human rights and well-being of European citizens

Freedom of speech, free access to alternative sources of information, transparent decision-making processes, fair elections, and inclusive, thorough debate on alternative solutions to issues are central to the functioning of our democratic societies. AI could be a powerful tool in strengthening the functioning and robustness of our democracies. Challenges to realising the positive potential include: 1) ensuring transparent decision-making processes. Information-sharing is an essential part of creating transparency. The question is what role AI-based technologies would play in information generation and sharing in the future. For information to support a trusting relationship it must be correct and not misleading information. The rise of Social Media as a source of information sharing already influences how politics is done, and it has introduced an increased need for controlling and checking information. The hard question is who decides what information is misinformation, and what is good information? Social Media tools can also be used to short-circuit democratic debate, and as tools for abuse and manipulation, 2) Very few private actors already control important channels of information, and in some cases decide on the limits of the freedom of speech. How will the balance between public and private actors develop in the future when it comes to digital spaces of information and debate?

We recommend increased focus on the quality and function of digital spaces supporting public debate. National and European debate and deliberations could be measures to nurture and grow well-functioning, accessible, inclusive, and non-discriminatory digital spaces for democratic debate and decision-making.

[12] Ibid.
[13] See e.g. the H2020 EU project CIMULACT.

Human oversight should be complimented by human insight and explainability

Requiring human oversight is an important aspect of the white paper. But oversight cannot stand alone, it needs to be supplemented with human insight. Without this AI becomes an opaque black box producing inscrutable decision suggestions. Without insight into the basis for the suggestion, the person acting on the recommendations from an AI cannot interpret the suggestions and their motivations properly, and thus cannot make proper use of AI for decision support. While this is problematic, since making use of advice requires understanding the basis for the advice, it is also problematic for other reasons.

First, it increases the risk of uncritically accepting or favouring the suggestion made by the system while ignoring contradictory evidence, something also known as automation bias. The consequence of automation bias is that even if there is human oversight, it is mostly so by name, not by nature. Automation bias is something that needs to be addressed by developing procedures for ensuring that AI systems do in fact deliver decision support and not de facto deliver decisions.

Secondly, without any understanding of how AI works and derives its conclusions, it is very difficult to determine that it is not functioning properly. For this reason, system errors risk going undetected for a very long time, possibly having substantial adverse consequences, be they social, material, economic or otherwise. Along the same line, bias in the system can go undetected by those working with it for much longer than necessary.

Thirdly, access to redress becomes arduous because the person making the decision does not fully understand the basis the decision is made on. Cathy O'Neil[14] presents a very good example of this: a schoolteacher is let go because she fared poorly in an assessment carried out by an algorithmic decision support system, despite being popular with the pupils and parents and not doing worse than other teachers at the face of it. No one was able to explain to the teacher why she had been let go, because the people who made decision based on the algorithmic suggestion, simply did not know why the system suggested that she had performed poorly.

Thus, for human oversight to be meaningful, human oversight needs to be supplemented by human insight, so the operator of the system can engage with the system and make decisions with support from it on an informed basis. To this end, it is necessary to ensure that those working with and making decisions with support from AI systems are trained to understand the system and the weighing of input as well as how the analysis is arrived at. Especially for systems which have a direct impact on peoples' lives.

Asking citizens to blindly have confidence in a decision reached by AI will not inspire trust, quite the contrary. Trust is inspired by being transparent about a decision-making process, including that being carried out by an AI. Thus, it should be a requirement that AIs used to make decisions that affects people directly should be explainable to the people who are affected by the decision. This is also an important element in ensuring the right to redress and to challenge decisions made by or with support from AI systems. In addition, this was a very clear message from citizens: they want to be able to understand the rationale behind an AI based assessment or decision that affects them.

[14] Cathy O'Neil (2016): Weapons of Math Destruction.

This is strongly linked with the issue of informed consent, which is almost not taken up in the white paper. However, discussing and finding solutions for informed consent is pressing for AI systems and application given the potential uses of AI and their potential privacy challenging capacities. Take e.g. systems for profiling, particularly predictive profiling. While the GDPR and other sector specific regulations address informed consent, it is nonetheless pressing that this is emphasized and strengthened with actionable tools in the regulation of AI. Given the potential uses of AI and the widespread implications it can have for democracy, privacy and fundamental rights, it is pressing that a European legislation on AI specifies this requirement, and that efforts are made to develop tools that can help obtain actually informed consent.

The importance of strengthening informed consent is further emphasized by citizens generally not feeling that they have control over their data, while at the same time being concerned about what it can tell about them. Empowering citizens to take control of their data is pivotal in creating public support for AI.

This project/research has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2), and the Specific Grant Agreement No. 945539 (Human Brain Project SGA3).