Response to the Consultation on the White Paper on Artificial Intelligence - A European Approach

Connecting Ecosystem of excellence and Ecosystem of trust: A concrete proposal

Authors:
Olga Afanasjeva
Marek Havrda, PhD
Will MIllership
The authors are affiliated to GoodAI Research, a Prague-base AI research start-up

AI is changing the world economy and our societies. It is important to realise that we are at the very beginning of a profound transformation. One possible way to picture this transformation is to imagine that each company as well as an individual has their own AI-powered assistant which significantly augments their abilities. In addition, more and more products and services will include AI-powered features which will allow them to "make decisions" previously reserved for people, including already today very personal choices about what we watch, what we listen to, or what we read. The transformation will bring about benefits in areas such as health, safety, or farming. On the other hand, AI may bring about new unintended impacts such as possibly new types of manipulation or addiction and new types of market failure. These new types of negative impacts may possibly require regulatory intervention in the near future.

As stated in the Communication Building Trust in Human-Centric Artificial Intelligence (COM/2019/168 final, further referred as 'Communication'): *AI is not an end in itself, but a tool that has to serve people with the ultimate aim of increasing human well-being* and AI applications have to adhere to ethical principles based on the European values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights. The Communication clearly states that unintended harms need to be avoided. In addition, the identification and assessment of potential negative impacts should be facilitated by the use of impact assessment which should be proportionate to the risks of a given AI system.

Building on the Communication and our experience of AI developers we propose a concrete way to effectively **connect the Ecosystem of excellence and the Ecosystem of trust.** For any regulatory intervention to be effective, it needs to be based on a sufficient understanding of the risks related to negative impacts and their root causes. Before putting in place any soft law and even more importantly any binding regulation the insights about the actual and potential impacts need to be gathered. **Therefore, the Commission needs to significantly increase its ability to monitor, assess and understand the impacts of AI in a multitude of concrete use cases at the level of individual users and at the level of society as a whole as outlined below. This could be facilitated by integrating impact monitoring and assessment systems within world reference testing centres mentioned in Action 2 of the White Paper.** In other words, apart from testing technical aspects including technical robustness and safety,

data processing, algorithmic bias, and explainability, the reference testing centres would allow for monitoring and assessment of impacts of AI applications.

First, in line with the Communication the Commission should develop a monitoring approach that would allow for the **assessment of the implementation of AI on well-being**. This is crucial for both identifying unintended consequences as well as the assessment of preliminarily identified risks. A potential way forward is to use existing standards on well-being impact assessment of AI systems and also human rights impact assessment. These approaches might utilise the 7010-2020 IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being which includes the detailed description of the related iterative process including consultation with users and other relevant stakeholders. The work of the OECD on well-being and New Zealand on the Living Standards Framework may also serve as inspiration for deriving relevant sets of well-being metrics for various AI use cases. The monitoring and assessment may also incorporate impacts such as new types of addiction in particular in children and increasing loneliness (not only of the elderly) and on individual health, both physical and mental. For example, TikTok, the most downloaded social app in the world used for creating and sharing very short (lasting 15 seconds or less) videos, is built around AI which generates the entire individualized feed based on quick analysis of viewing behaviour and often leading to binge-watching video content. The user may select preferred sources and topics, but concrete content in the stream is generated by AI. A survey of 60 000 families in the US, UK, and Spain found that children 4-15-years old spent during the covid-19 quarantine about 82 minutes watching TikTok videos daily (and e.g. in Spain TitTok is used by almost 40% of surveyed children). TikTok has been downloaded over 2 billion times and is available in 155 countries and 75 languages.

Second, **at the level of individuals, special attention should be paid to human agency.** The Commission needs to develop sound approaches and methodologies for monitoring the impacts regarding fundamental human rights and freedoms including potential emotional and cognitive manipulation enabled by AI systems, i.e. impacts directly related to human agency and autonomy. One of the use cases where AI may contribute to the reduction of human agency (in particular in children) is the successful deployment of AI to increase engagement on social media. This is in line with the Communication which stipulates that the main focus should not be only on the overall wellbeing of the user, but *AI systems should support individuals in making better, more informed choices in accordance with their goals.* The systems should support human agency and fundamental rights, and not decrease or misguide human autonomy. **The Commission should develop new metrics and new approaches related to human agency when AI systems are deployed and new metrics for monitoring whether intentionally or unintentionally the AI systems are abusing human biases leading to manipulation.** In addition, a specific measuring approach may need to be developed also regarding other key requirements mentioned in the Communication, in particular transparency and fairness. All these metrics would be integrated and analysed by the impact monitoring and assessment systems.

Third, apart from the impacts at the level of users and other involved individuals and other stakeholders, we need to understand better **what impacts AI deployment may have at a societal level, in particular on social cohesion and democracy**. The Communication stresses the need to consider the impact of AI systems on society as a whole particularly in situations relating to the democratic process, including opinion-formation, political decision-making, or electoral contexts. Nevertheless, the White Paper does not pay sufficient attention to this crucial level of impact. For example a report "The Future of Political Campaigning." published DEMOS outlines several ways AI can be (ab)used in political campaigns including using AI to automatically generate content and micro-targeting together with 'psychographic' techniques enabled by deep learning. Again, new relevant metrics will need to be developed at the societal level in order to assess both intended and unintended impacts of various implementations of AI. The efforts need to be at least co-created by public actors, such as businesses, even the largest online platforms, are not well-positioned to deal with these types of impacts in part due to the potential conflict of interest due to prevailing business models.

In terms of priorities based on various types of the AI applications or sectors, these new assessment methods should first be developed in detail for the potentially most harmful and pervasive applications at the level of individuals such as biometric recognition systems, personal recommender systems, and other assistive technologies which may (ab)use AI for emotional and cognitive manipulation, and systems that systematically gather and assess personal data from multiple sources. Several companies including Microsoft and Google are already using internal systems for Human Rights impact assessment which may serve as inspiration. At a level of societal impacts, the attention needs to be paid to those which may significantly disrupt democratic processes.

Developing these kinds of monitoring and assessment systems seems to be beyond the traditional abilities of individual companies. Public (or not-for-profit or joint public-private) actors may be better placed to develop them and offer them for free and voluntary use by businesses, large and SMEs, and other actors considering the deployment of AI. The monitoring and assessment system would assist the deploying entity to specify which data needs to be gathered and analysed in order to understand related wellbeing and other significant impacts, both intended and unintended. It may be worthy to assess data from relevant reference non-implementation sites or populations in line with A/B testing approaches or Randomised Controlled Trials in order t**o improve assignment of causality between the AI system implementation and observed impacts based on data.**

The main motivation for the private sector to voluntarily take advantage of such monitoring systems would be related to reputation and general risks management. Among the motivation for public actors would be to gain a much better understanding of potential impacts and in the future creating an environment (sand-box) for testing potential regulatory action. **Ensuring the security of data and the safeguarding of intellectual property rights are among the main prerequisites of the successful deployment of such an impact monitoring system.** The

results of monitoring and assessment should be at first available to the companies with confidential reporting mechanisms in case negative impacts are identified. It is **imperative to build impact monitoring systems on the basis of high trust**. Institutions with high credibility and trust such as the OECD, IEEE, or consumer protection organisations may be among the entities to be involved among the partners developing and operating such systems.

Taking into account that the Commission aims to target and limit regulatory actions to *clearly identified problems* as stipulated by in the White Paper (p. 10) much more efforts need to be devoted to the problems of monitoring and identification. Integrating impact monitoring and assessment systems within proposed reference testing centres would allow us to better identify and assess various trade-offs between the ethical requirements in terms of impacts which will be in practice often unavoidable. Finally, **testing regulatory remedies in a near-real-world setting** (as suggested by GoodAI in the 2018 article “Building on the idea of an ethical framework for a Good AI Society”) **would significantly increase the probability of the success of regulatory intervention attaining its objective.** Regulating AI requires 21st century approaches, i.e. building policy interventions not only on intuition about causal effects but as much as possible on data. The proposed agile approach to impact assessment would also considerably **boost the responsiveness of regulatory action** which seems to be imperative in the realm of fast-evolving research and deployment of artificial intelligence, especially considering its potential far-reaching impacts.

References:

"20 TikTok Stats For Marketers: Tiktok Demographics, Statistics, & Key Data". 2020. Mediakix. https://mediakix.com/blog/top-tik-tok-statistics-demographics/.

“Apps and Digital Natives: The New Normal Connected More than Ever,” Qustodio 2020 annual report on children’s digital habits. 2020

Bartlett, Jamie, Josh Smith, and Rose Acton. “The Future of Political Campaigning.” DEMOS, 2018. Accessed June 14, 2020. https://qweb.cdn.prismic.io/qweb/e59c2e0f-ef4f-4598-b330-10c430e2ec71_Qustodio+2020+Annual+Report+on+Children%27s+Digital+Habits.pdf.https://demosuk.wpengine.com/wp-content/uploads/2018/07/The-Future-of-Political-Campaigning.pdf

“Building on the Idea of an Ethical Framework for a Good AI Society | GoodAI.” GoodAI, December 21, 2018. https://www.goodai.com/building-on-the-idea-of-an-ethical-framework-for-a-good-ai-society/.

COMMUNICATION Building Trust in Human-Centric Artificial Intelligence, COM/2019/168 final

Davis, Jason. "The Tiktok Strategy: Using AI Platforms To Take Over The World". 2019. INSEAD Knowledge. https://knowledge.insead.edu/entrepreneurship/the-tiktok-strategy-using-ai-platforms-to-take-over-the-world-11776