Comments by Johann Čas on the European Commissions' consultation on the inception impact assessment as part of the initiative "Artificial intelligence – ethical and legal requirements".

Johann Čas is a senior researcher at the Institute of Technology Assessment of the Austrian Academy of Sciences. Among other things he is co-author of the study "When algorithms decide in our place: the challenges of artificial intelligence in Switzerland" (https://www.oeaw.ac.at/en/ita/projects/the-social-effects-of-artificial-intelligence/) (the main text is only available in German) and a team member of the H2020 PANELFIT project consortium (https://www.panelfit.eu/). All opinions and views expressed are personal.

Welcoming the opportunity of providing feedback on this legislative initiative I would like to provide the following comments and suggestions:

A. Context, Problem definition and Subsidiarity Check

The introduction appears to be based on overoptimistic perspective. Whereas the positive potentials certainly exist it should be also mentioned that they are also accompanied by corresponding risks, that their materialization depends on the creation of suitable framework conditions, which must also include a fair distribution of the possible benefits. A more realistic and sober assessment of opportunities and risks would be more helpful as a basis for deciding on regulatory options. To explain this recommendation in more detail, here an excerpt from a commentary I wrote during the consultations on the Ethics guidelines for trustworthy AI of the High-Level Expert Group on Artificial Intelligence (HLEG):

"Last but not least, the working document [draft Ethics guidelines for trustworthy AI] appears to overestimate the actual and potentially positive contributions of AI to solve the grand challenges our world is facing, missing to provide evidence for this positive overall evaluation. Whereas large and important positive potentials can be envisaged, past and current use of AI does not appear to support this judgement. Taking increasing economic inequality as an example, AI has rather contributed to it – e.g. in form of a key enabling technology of high-frequency trading on financial markets - but I'm not aware of making serious attempts to use AI to resolve imbalances on labour markets. On the political level, AI is rather threatening civil liberties and democratic systems than empowering citizens. On a global level, AI is rather supporting the establishment of worldwide monopolies than empowering consumers. Data based businesses possess unprecedented economic capacities, unparalleled political influence, powers to shape the results of democratic votes, unique possibilities to influence or to manipulate individuals in the information they receive or decisions they take. By disproportionately stressing the potential positive impacts and neglecting the already materialised threats the working document in the current form might contribute to an inappropriate reliance on AI when tackling urgent problems of the EU and the world. By missing to mention these dangers and threats it also misses to analyse them and consequently to develop measures and policies to counter them. This leads back to the ~~first~~ [next] critical comment: we are primarily not in need of a trustworthy technology but of making good use of opportunities offered by technology in the human interest and keeping human agency."

The term " trustworthy AI" should be avoided, although it is used in many key documents. This term suggests that, provided certain conditions are met, this technology can be used without further concern or human supervision. Thus, the use of this pair of terms contradicts one of the key requirements of " trustworthy AI", namely human agency and oversight. To explain this

recommendation in more detail, here another excerpt from the commentary that I wrote during the consultations on the Ethics guidelines for trustworthy AI of the High-Level Expert Group on Artificial Intelligence (HLEG):

“The first critical comment is related to the selected title as such: there are only a few technologies imaginable that can be regarded as trustworthy on their own. In the case of AI, the naming of “ethics for trustworthy AI” is in appropriate and misleading for several reasons. Depending on the concrete AI technology in mind, the results produced by these technologies are at least prone to statistical errors, some also show completely unexplainable (and unpredictable) behaviour. The labelling as guidelines for trustworthy AI contains at least implicitly the message that trust in these technologies is in principle justified as long as the developed guidelines are respected, neglecting the fact that it is the use of technology only, which could deserve this marking. In the case of AI this comment, which might be regarded as a linguistic sophistry, is doubly important. AI is threatening human autonomy and agency, as acknowledged in the working document; neglecting this fact in the very title is additionally endangering human agency. At least a renaming in the form of “ethics guidelines for trustworthy use of AI” or something similar should be considered. Otherwise the guidelines could become self-contradicting to one of the core principles mentioned in the document.”

The statement that “Just like for actions and decisions taken by humans, the use of AI to either directly take decisions or to support decision-making may lead to violations of fundamental rights, as guaranteed by and implemented in EU law.” negates fundamental existing ethical concerns and legal restrictions on automated individual decision making in the European Union. By including direct decisions by AI as a possibility, it pre-empts debates on this issue and the extent to which we should grant such powers to AI.

B. Objectives and Policy options

AI is a very potent technology, affecting almost all areas of life and the economy, which brings with it great opportunities but correspondingly high risks. Accordingly, option 4 should be chosen and designed, as none of the previous options adequately addresses the positive and negative potentials.

Option 4 does not exclude corresponding initiatives by industry or voluntary forms of labelling. However, such activities can by no means be sufficient, but must be embedded in an appropriate regulatory framework.

The regulatory framework should be designed gradually according to the magnitude of risks, in analogy to the risk pyramid of the German Data Ethics Commission. AI applications that do not or only indirectly affect humans do not require specific AI regulations. This includes, for example, AI applications for the analysis of large amounts of data in astronomy or physics, or applications in the field of industrial automation. However, the second example does require political measures, more on this below.

For all other AI applications, appropriate precautions must be taken according to the risks involved. For example, recommendation systems or filter algorithms in social networks also need to meet transparency requirements. The restriction to high risk applications proposed in Option 3B is totally insufficient. First, it is not clear why two risk conditions have to be fulfilled simultaneously for AI applications to fall into this category of need of regulation. Second, such classification criteria hinder dynamic adaptations to new applications and associated risks. Moreover, high-risk applications should always be questioned in principle, taking into account not only the effects on individuals but also societal consequences (see below).

In principle, a broad, technology-neutral definition of AI is preferable.

C. Preliminary Assessment of Expected Impacts

In the section on possible economic impacts, a comparison of compliance costs caused by regulation with the benefits of AI is addressed as one basis for decision making on legal obligations. The potential costs of non-compliance are, however, almost completely neglected. It is in principle questionable whether calculations which compare the protection of fundamental rights with economic costs should be permissible in democratic societies.

If such comparisons are made, however, at least the costs of non-compliance should be considered accordingly. This includes not only the economic costs, for example in the form of external effects or costs caused for the industry by acceptance problems, but also the costs for society, social coexistence and democracy. If the broad and deep impacts of AI applications are realistic, which are always emphasised in the case of positive effects, they should also be taken into account in the case of possible negative effects. These include in particular effects on the social and political climate that have already occurred or are becoming apparent. The principle of precaution requires that preventive effects be assessed and controlled in order to avoid future poisoning of social coexistence or overheating of the global political climate in analogy to the environmental pollution or global warming

The section on possible social impacts contains a statement in which two auxiliary verbs have been mixed up: "… while AI-enabled automation **may** replace some jobs, the use of AI **will** simultaneously lead to the creation of new jobs …", correctly it should read like this: "… while AI-enabled automation **will** replace some jobs, the use of AI **may** simultaneously lead to the creation of new jobs …",

Automation will in any case replace jobs as a direct effect; no company will invest in new technologies unless the investment in technology is expected to yield more than compensating savings in labour costs. The extent to which these lost jobs can be partly or fully compensated by newly created jobs depends on a number of external factors that cannot be predicted. A trustworthy policy must ensure and guarantee that, whatever the actual development will be, the benefits of automation are fairly distributed and that this does not translate into high and persistent unemployment. This demand goes far beyond the direct regulation of AI applications, but it is indispensable for a broad acceptance and therefore also the fullest possible use of the rationalisation potential of AI if the social coherence and political stability of the European Union is not to be jeopardised even more.

D. Evidence Base, Data collection and Better Regulation Instruments

As the response to this initiative also shows, online consultation is not sufficient for a broad participation of citizens. In view of the great importance and the many changes that AI brings to citizens of the European Union, targeted citizen participation activities involving all Member States would be appropriate and should be considered.