Non paper NL Ministry of Interior and Kingdom relations & Ministry of Economic Affairs and Climate on the draft EU ethics guidelines for trustworthy Artificial Intelligence (AI)
31 January 2019

General impression

- ***Suppor***t: NL supports the initiative of the European Commission and work of the High-level Expert Group on Artificial Intelligence (HLEG AI) which is a good -diverse-representation of businesses, NGO's and social partners. The draft guidelines strike a good balance between risks and opportunities, AI should be trustworthy and respecting fundamental human rights in applicable, responsible manner. From the Member States' perspective these guidelines should be an element of the different national AI strategies. Which, in turn, are in line with the EU Coordinated action plan.
- ***Tone of the document:*** we should have more emphasis on how to stimulate good AI, rather than control bad AI (now only the conclusion contains a more positive tone). We want to achieve trustworthy AI, that will be seizing the opportunities and creating a leading position for Europe on AI.
- ***Purpose, line of reasoning, six suggestions*****:**
    - The goal and purpose of the draft guidelines are not completely clear. How can we use the guidelines? If it's not going to be a directive, what can we do with it? ***Suggestion 1***: we should add more European global thought leadership to the guidelines and ambition: a statement that the EU wants to become AI leader. Europe will only be able to set standards if it has something to say in terms of innovation and adoption and then it is more logical to incorporate it in all national AI strategies of Member States. The global reaction to the EU's General Data Protection Regulation (GDPR) has shown a strong framework which can shape global markets and strengthen the EU economy.
    - If the EU wants to be a leader, it is good to have our own principles. A lot of the principles based in other continents are developed in, for example, the BIO tech industry (page 6 of the draft guidelines).
    - ***Suggestion 2:*** We should be focused on our own European principles, including human rights and a human centric approach as mentioned in the guidance.
    - In order to be leading, the EU must find a way to approach AI from a global perspective. What are we going to do with AI products in the EU that would not be compatible with our European framework, and AI products that may be applied in line with our framework, but that are trained or developed in a way that are not compatible?
    - ***Suggestion 3***: Algorithms: transparency and explainability are of great importance to gain more trust from citizens and increase accountability. Algorithms are not in the glossary or specifically mentioned, we would suggest an extra paragraph. Black box algorithms have our special attention.
    - ***Suggestion 4*****:** Make the guidelines more globally oriented: it is therefore important to work together as Europe in other global fora, for example the OECD, ITU and G20.

- The roles of private industry and public sector need to be further determined. How do we make the guidelines operational?
- ***Suggestion 5*:** It can be made practical through experiments. Trial and error, ethics in, by and for design is creating a new situation with trustworthy AI. It also tackles the dynamic and pragmatic reality. Standards and *self-evaluation* can play an important role *self-regulation* will only work if we have smart evaluation or ISO/audit systems to make this work. That will speed up our own learning curve on trustworthy AI and foster the willingness of consumers and institutions to adopt AI-applications, which could offer opportunities for Europe.
- We would suggest to refer to existing European Commission Better Regulation documents or OECD documents, like the OECD Due Diligence Guidance for responsible Business Conduct.
- ***Suggestion 6:*** Three elements that could be used as the key organizing principle for the public and private sector: ethical purpose of the guidelines, technically robust guidelines and a leading role for Europe at a global stage, this could be reflected in the title.

Specific points, in addition to the general impression

- Page iv glossary, definition of AI: systems that act in the physical or digital world, by perceiving, interpreting, reasoning and deciding." This definition assigns technology in a way that could be misleading: nobody talks about a thermometer "perceiving" temperature, although it "decides" very clever "how many degrees it is" based on "what it perceives". Just like a thermometer an autonomous car, robot or character recognition device does not perceive anything. The machines are configured (not "trained") to map input on a predefined output: turn the steering wheel, initiate or stop a process or output something. So, when an AI device "does not understand me", this means it cannot match the input with an output. This happens either because the input was never presented before or because there is not a suitable output.
- Page 1, Trustworthy AI, a key element. We would suggest to add numbers: how many people want trustworthy AI and an increase in research on the importance of trust.
- Page 2, role of AI Ethics. It is a starting point, not the finish line. In this section the purpose and role of the guidance can be added. Also the importance of experiments as well as trial and error. It could be made more vivid by turning it into a tool, using best practices and results of previous cases.
- Page 3, Scope of the Guidelines: if you describe the scope, describe what it is rather than what it is not.
- Page 3, A Framework for Trustworthy AI. realisation of trustworthy AI. We would suggest to include human rights and make our own (European) principles. Use cases: make it dynamic, trial and error and lessons learned.

Greater emphasis for selected issues:

- Validation is important: we should suggest to further define what is tested and what is not, as well as how the testing will be done.
- Good graphics: the greater the feedback loops, the fewer testing requirements
- The four user cases for corporate and public AI mentioned in the document should be presented at the beginning of the document (it's a good delivery of the HLEG AI and follow up).
- The tensions concerning transparency, i.e. between transparency and innovation (when making public business secrets) and transparency/explicability and added value of AI (when demanding transparency or explainability means that specific types of AI, like deep learning neural networks, cannot be used as they are inherently opaque). This is very briefly addressed in the summary and on page 23 (i.e. gaming the system), but is not further illustrated.

Ethical principles (page 8 to 10):

- Ethics by design should – from a technical perspective- be further examined with ethics by the adaptive system and ethics by behaviour.
- The term 'explicability' is problematic (if required in common language), explainability could be an alternative We would suggest to highlight the 3 terms: **describe, inspect and reproduce** (= which de-facto translates into auditability)
- Related to this, some concepts seem to be used interchangeably: explicability (p10, but is not used in the rest of the doc), traceability (p20, which might mean the same, but only in a specific case), explainability (p21) are all used, but seem to cover more or less the same idea. We would suggest to Please clarify these concepts and reduce the use of different concepts throughout the document.
- Problematic: training data also falls under GDPR (for example: how do we define consent for training data) Is there a way to exempt specific consent requirements (for example in public health sector)
- Standards: EU first then adoption elsewhere, or global from the start
- Accountability governance: suggested title for function in organizations: Data stewards? (already used for data analytics governance, and gained traction recently. Building on what is accepted might help, rather than coming up with new ideas.
- Human oversight: state the need to add resources to organize frequent conferences for (international) knowledge exchanges – EU must frequently showcase what is being taught/applied.
- R&D: traceability and explainability: **how** decisions come about (**not why**) – need for more research (XAI): state that EU should work on incentives from public agencies to stimulate this type of R&D so that EU becomes leader in XAI as a growing academic field.
- New categories: when working on possible codes of conduct: introduce specific types of AI (for example: decision-support is different from autonomous systems, lethal force is different from consumer products)

- Suggestion to the European Commission to establish EU and/or national awards for best practices from private and public entities.
- Suggestion to establish academic programmes to work on trustworthy AI (MA programmes for example) in close cooperation with the Member States, academia and businesses or, suggest a EU Erasmus type programme to help stimulate uptake across Europe

Specific additions:

- Failsafe shutdown is a good proposal; but we might need a European or international arbiter to keep record of when it was used, by whom. (this could be an additional item for non-technical methods section)
- Transparency: add the question on whether a warning sign for user might be useful whenever a personal ID is determined / used? Think of incentives not to go all the way at once, but to start with supervised AI (ethically easier stuff)
- "Common good" is very promising, but hardly elaborated: common good is interpreted (as elaborated in the report) in a narrow way as "contributing to a good life". That is a reduction and potentially harmful to ethical application of AI. Therefore "common good" by "well-being" as defined by the OECD
- Another non-technical method is the implementation of an AI Impact Assessment. In the NL we have published an AI Impact Assessment". This could be designed in line with the design of a privacy, data protection or human rights Impact Assessment. A strength of this approach is that it can properly take into account the specific context.