Translational model for Human Oversight measures

Article 14, paragraph 4 of the proposed AI Act states that *"The **Human oversight measures** shall enable the individuals to whom human oversight is assigned to do the following, as appropriate to the circumstances".* All the five measures described in the Proposal are interpreted below in three different layers: technical layer, socio-technical layer and governance layer.

*(a) fully **understand the capacities and limitations of the high-risk AI** system and be able to duly monitor its operation, so that **signs of anomalies, dysfunctions and unexpected performance** can be detected and addressed as soon as possible;*

1. Technical layer:

a. Engineers of the system will be able to accurately describe what the AI system can and cannot do by describing the bounds of the data sets and its limitations. Engineers will build in test-cases and code that can detect unexpected outputs and failures of the system. Engineers will create a reporting system for assigning unexpected behavior tickets to engineers to fix.

b. During the testing phases, engineers should also test with biased inputs to measure the threshold of the accuracy in the system's results/outputs. This can also assist developers to identify what a misuse of the system would look like. This can help to recheck the data elements classification from which the results are derived.

2. Socio-technical layer:

a. Educators and other professional educational staff that interact with the AI system should receive training on expected outputs, how they may detect unexpected behavior of the system, who to report unexpected outcomes to, and what non-AI systems to fall back on should the system behave in unexpected ways.

3. Governance layer:

a. The measure described as ***"fully understand** the capacities and limitations of the high-risk AI system"* could be implemented by using counterfactual explanations, which specify what circumstances would need to change to achieve a more desirable decision, in contrast to explanations that involve an attempt to outline the logic of algorithms. Counterfactual explanations attempt to address the **human interpretability** issues inherent in machine learning algorithms. Counterfactual explanations **do not require** individuals to understand any algorithms in order to extract a meaningful explanation. They are easy to understand and practically useful as they provide the circumstances that need to change to achieve a more desirable decision. (Gacutan & Selvadurai, 2020).

(b) remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system ***('automation bias')****, in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;*

1. **Technical layer:**

    a. Engineers should be made aware that the eventual goal of the system is not assessment, but identification of at-risk students for outside services to assist them in their education. It should be built into the system notifications for when someone is interacting with AI or when they are viewing a suggestion or claim made by an AI and not a human.

    b. When providing the recommendations for at-risk students, the system should be capable of explaining in simple language what are the factors that the AI considered in order to make the recommendation as appropriate to the circumstances and existing interpretability capabilities. Understandability of the explanations provided is another critical factor that engineers should be aware of when developing systems for sensitive user groups, such as students.

2. **Socio-technical layer:**

    a. Educators should be encouraged to push back on and question the decisions of the AI system. For instance, if an instructor feels that a student would have a better school experience with additional support, but the AI system has not identified that student for support, the instructor should feel free to offer that student support services.
    b. Educators and managers should come to an agreement about lines of accountability for decisions made or informed by algorithms to overcome issues related to blame avoidance or deference to algorithmic recommendations.

3. **Governance layer:**

    a. This measure requires that human oversight is to be ensured in such a way to enable the person assigned that task to be able to be aware of the potential of 'automation bias'. Automation bias occurs when an operator relies solely on automated recommendations without searching for disconfirming evidence. (Atchley, Smitherman, Simon and Tenhundfeld, 2020).

    b. The explicit recognition of this problem is valuable in itself. Yet, combatting it more effectively might necessitate additional safeguards, for instance by requiring the Education Division of the City of Helsinki to communicate how other available information or alternative outcomes were considered in reaching a decision. (Fink, 2021).

(c) be able to correctly ***interpret*** *the high-risk AI system's* ***output****, taking into account in particular the characteristics of the system and the interpretation tools and methods available*;

1. **Technical layer:**

   a. In the design of the system, the outputs should be immediately understandable even to those with no technical background. For instance, converting psychometric outputs to plain, easy to understand language. In the design of the dashboard, it should be made clear what data sources are being used in an AI system, the limitations of that data, and how an output is constructed from those inputs.

   b. The use of jargon and technical terminologies should be avoided in the system's explanation to make it understandable for any audience.

2. **Socio-technical layer:**

   a. Educators should be trained on expected outputs for on-track and at-risk students, and how those outputs translate into access to real world support systems for graduation success. For instance, in the case of different levels of at-risk identifications, what types of support would a teacher be able to call upon for different kinds of students?

3. **Governance layer:**

   a. The right to explanation envisaged in the European Union's 2018 General Data Protection Regulation (GDPR) allows an individual to seek 'meaningful information' about the 'logic' involved in making a decision, in circumstances where the decision was made solely using automated technologies and the decision produced legal effects concerning the individual or significantly affected them. Although the legal status of this right to explanation has been the subject of considerable debate, it is an important development towards human oversight, interpretability and explainability.

   b. Explanations of specific algorithmic decisions should allow the justification of a black-box model or decision to be debated and contested. Further, meaningful, critical dialogue must be achieved between user, developer, and model by ensuring explanations are contrastive, selective, and social. (Mittelstadt, B., Russell, C., & Wachter, S., 2019).

c. Participation—including the related requirement of information transparency—and accountability are inter-related principles that build on each other in the practice of human rights; it is only when people have the information and can participate in decisions that the designers and users of AI design, development and deployment can be held to account. (Fukuda-Parr & Gibbons, 2021).

(d) be able to decide, in any particular situation, ***not to use the high-risk AI*** *system or otherwise* ***disregard, override*** *or* ***reverse*** *the output of the high-risk AI system;*

1. **Technical layer:**

   a. Engineers should design a system that allows educators access to non-AI decision making data, so that users can fall back on their own data analysis should the system behave in unexpected ways. In the pre-implementation stage where engineers are training models, they should have access to the governance layer to explain why they believe a model may not work or cause unintended harm.

2. **Socio-technical layer:**

   a. Educators should still be able to make their own decisions about what students need and what additional supports would be best for their academic success. Educators should have access to engineers to express their concerns about when to disregard or override the suggestions of the AI system.

   b. [Should students, parents, and public sector managers also be able to make decisions about when not to use high-risk AI systems, for example around automated or semi-automated support, or dashboards?]

3. **Governance layer:**

   a. The Human-in-command (HIC) approach refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the **ability to override a decision made by a system.**

b. Human oversight needs to be meaningful in the sense that the overseer should have the authority and competence to change the decision (Brkan, 2019).

c. Art. 22, par. 3, GDPR: "*The data controller shall implement suitable **measures** to safeguard the data subject's rights and freedoms and legitimate interests, **at least the right to obtain human intervention** on the part of the controller, to express his or her point of view and to **contest the decision."***

d. This human oversight measure observes the fundamental right to an effective remedy (Art. 47 EU Charter of Fundamental Rights).

*(e) be able to **intervene** on the operation of the high-risk AI system or interrupt the system through a **"stop" button** or a similar procedure.*

1. Technical layer:

a. The system should be designed such that the AI functionality can be completely pulled from AI-HOKS and not interrupt the student data stream. While the AI system is offline for any reason, educators and students should still have access to their VET data. This can be designed by integrating the AI system in a modular way into AI-HOKS to not disrupt the service.
b. During the occurrence of any adverse operation, the AI recommendations or suggestions should be turned off. Users should still have access to the service where they can interact such as, viewing their grades, past/current/upcoming courses etc. Whereas the recommendation AI should be deactivated and directed to the testing team for error identification.

2. Socio-technical layer:

a. In the event of a system stop, educators should be trained on how to access the data in other sources or databases should AI-HOKS go down. If a user believes that the system is acting in a way causing widespread harm, then they should have access to an escalating system that alerts engineers and governance stakeholders.
b. When the recommendations/suggestions provided by AI are halted because of anomalies, students should still be able to see their grades, course lists, and other basic course/personal information without interruption. AI used here should only enable the students with additional support rather than taking away the basic access to the service.
c. [How will students, parents, and managers interact with the data?]

3. Governance layer:

a. This measure highlights the need and importance of human autonomy when applying oversight to AI systems. The oversee should have the capacity, autonomy, and power to be able to intervene and stop the AI system.

b. It is really important to address the existing power relations when human oversight is being deployed. If individuals to whom human oversight is assigned do not have enough autonomy to effectively intervene in AI operations, the purpose and efficacy of oversight measures will be compromised and their impact will be undermined.

References:

Brkan, M. (2019). Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond. *International Journal of Law and Information Technology*, *27*(2), 91–121. https://doi.org/10.1093/ijlit/eay017

Davis, J., Atchley, A., Smitherman, H., Simon, H., & Tenhundfeld, N. (2020). Measuring Automation Bias and Complacency in an X-Ray Screening Task. In *2020 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 1–5). IEEE. https://doi.org/10.1109/SIEDS49339.2020.9106670

Fink, M. (2021). The EU Artificial Intelligence Act and Access to Justice. *EU Law Live*.

Fukuda‐Parr, S., & Gibbons, E. (2021). Emerging Consensus on 'Ethical AI': Human Rights Critique of Stakeholder Guidelines. *Global Policy.* Advance online publication. https://doi.org/10.1111/1758-5899.12965

Gacutan, J., & Selvadurai, N. (2020). A statutory right to explanation for decisions generated using artificial intelligence. *International Journal of Law and Information Technology*, *28*(3), 193–216. https://doi.org/10.1093/ijlit/eaaa016

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. In Association for Computing Machinery (Chair), *FAT* '19: Conference on Fairness, Accountability, and Transparency,* Atlanta GA USA. Retrieved from https://doi.org/10.1145/3287560.3287574