AI Ethics Systemic Translational Matrix for AI and Learning Analytics at Vocational Education and Training (VET) in Helsinki

Ethical requirement (High-Level Expert Group on AI, 2020)	**Other meanings** (Fjeld, Achten, Hilligoss, Nagy, and Srikumar, 2020)	**Concrete risks and harms addressed** (risk-based approach and human rights-based approach)	**Applied to Educational Sector** (Vincent-Lancrin, S. and R. van der Vliesm, 2020; Slade, Sharon & Tait, Alan, 2019; Miao, Holmes, Huang, Zhang, 2021)	**Translated to the use case**	**Technical interpretation** (European Commission. Directorate General for Communications Networks, Content and Technology., 2020, Weyns, 2020, Gotterbarn et al., 1999, Thomson & Schmoldt, 2001)	**How to involve stakeholders?**
1. Human Agency and Oversight	"human review of automated decision," "ability to opt out of automated decision," and "human control of technology."	Human out-of-the-loop. Decisions are not explainable. There is no obvious accountable party. Automatic decisions cannot be contested by the person they impact, nor can they be modified in exceptional circumstances by public administrators.	Student agency and responsibility (Slade & Tait, 2019).	Although it is clear that there is an asymmetrical power-relationship between institutions and students, proactive engagement at least seeks to treat students as equal participants in the uses of their data. In this way, students can be more actively involved in helping the institution to design and shape interventions that will support them. (Slade & Tait, 2019).	Users have control about how their data enters the system and understands the lifecycle of their data. Additionally, an understanding of how AI processes their data. It is clear to the user whether they are interacting with a human or AI. How is human oversight implemented into the design?	Use participatory tools that could elicit information about whether users feel reductions in their agency or situations where the technology exhibits undue control. Ice breaker exercises, Contextualizing with scenarios (Scenario-based approach), embodied participatory methods (e.g., walkthrough or media go-along) (Malinverni et al., 2019; Light et al., 2018; Jørgensen, 2016)
2. Technical Robustness, Accuracy and Safety	"safety," "security," "security by design," and "predictability".	Risks of a negative impact because of unreliable or low quality decisions. Risks of Cyber-attacks (e.g., ransomware, denial of service, data breach). Accuracy amounts to understanding performance, identifying the sources of error and the limitations of a solution and considering the quality and reliability of the decisions, as well as their direct societal impact. (Unceta, Nin and Pujol, 2020)	Students seem to be quite positive about the possibilities of learning analytics but are also concerned about the safety and usage of their personal data. (Nevaranta, Lempinen, & Kaila, 2020).	Inaccurate data could lead to inappropriate evaluations of progress or recommended supports. Security breaches could release personal information that could lead to harms (such as stigmatization or discrimination).	Relevant safety measures are built into the codebase and backend. There are protocols to deal with any potential breaches. Cyber-security measures have been taken in accordance with EU law. The technology is accurate and there are mechanics to monitor or improve accuracy.	Use participatory tools that could elicit the most relevant information about where risks of data inaccuracy or data security could lead to the greatest harms. Interviews, focus group discussion, Online surveys, Observation
3. Privacy and Data Governance	"privacy by design," "consent," "control over the use of data," "ability to restrict data processing," "right to rectification," "right to erasure."	Risks to the **right to private life** (Art. 7) and the **right to the protection of personal data** (Art. 8, EU Charter of Fundamental Rights). This dimension accounts for risks in three forms: **reidentification risk, data linkage risk**, and **sensible attribute inference risk**. Reidentification considers the probability of identifying an individual in the training set. Data linkage concerns the probability of being able of linking/joining two different datasets. Sensible data inference concerns the problem of using a ML system to infer protected information. This risk involves the leakage of sensible information through other attributes (Unceta, Nin and Pujol, 2020).	The collection and storage of data create new risks for privacy of students. Beyond the "Big Brother" fears that are common to all sectors of society, additional concerns related to privacy and AI in education usually are at least twofold. Families are concerned that education institutions or even employers may use "old" data to make decisions, which raises the question of how long and which data could be stored and retrieved to make some decisions. A second question relates to the possible use of the data for commercial purposes in a sphere where commercial interests are often excluded. (Vincent-Lancrin and Van der Vlies, 2020).	Students should have some input to determine which data can be collected, how that data can be used, who is able to access it, and for what purposes (Prinsloo & Slade, 2017). CoH should grant students the ability to correct and/or add context to their raw data, and to review and make a case for choices which appear to be limited as a result of a learning analytics application.	Users can pull their data from the system at any time, and there are protocols in place to implement this. Oversight mechanisms for data processing (including limiting access to qualified personnel, mechanisms for logging data access and making modifications). Measures to achieve privacy-by design and default (e.g. encryption, pseudonymisation, aggregation, anonymisation). Establish mechanisms that allow flagging issues related to privacy concerning the AI system.	Use participatory tools that could elicit the most relevant information about individual privacy preferences (both expressed and enacted). Identifying and prioritizing benefits and privacy risks of using LA through **participatory design**, by applying **Nominal Group Technique (NGT) with students**, exploring their perceptions of privacy protection and the use their data. In the case of privacy risks, understanding the perspective of the data subjects (eg, students) is critical because the privacy-enhancing design options are predominantly for their benefit. Moreover, because individuals' privacy preferences often conflict with their actual behaviors (ie, the privacy paradox), understanding which **privacy-enhancing design options** are most critical cannot be accomplished without proper engagement. This engagement not only leads to better software, but also participatory software design can importantly increase transparency and trust in AI. (Giannouchos et al., 2021).
4. Transparency	"transparency," "explainability," "open source data and algorithms," "open government procurement," "right to information," "notification when interacting with an AI," "notification when AI makes a decision about an individual," and "regular reporting."	A lack of transparency about how tools work could lead to a decline in public trust (Janssen et al., 2020), or to system avoidance behaviours (Brayne, 2014).	Transparency can be considered as one of the most crucial factors regarding the acceptance of learning analytics systems: This involves disclosing information about the collected data, its purpose, the underlying algorithms, the people who receive access to the data, and the analyses derived from them, as well as the amount of time the data will be stored and its degree of de-identification (Pardo and Siemens, 2014).	If teachers do not know if or how the tool will be used to manage their performance, they may resist its use (Jakobsen et al., 2018; Mumtaz, 2000).	Building warnings for the user when they are interacting with AI, making data open source where possible, making code open source where possible, transparent data cleaning processes. The decisions the AI makes are understanble and traceable in such a way that the average user can understand how their data is processed. The purpose of the AI is communicated clearly to users.	Use participatory tools that could elicit information about whether users feel there is sufficient transparency or to help identify where they would like to see more information available. Storyline workshops, Contextulizing with scenarios (Scenario-based approach), web-based tools to facilitate participation and engagement (Viale Pereira et al., 2017).
5. Diversity, Non-discrimination and Fairness	"non-discrimination and the prevention of bias," "representative and high-quality data," "fairness," "equality," "inclusiveness in impact," and "inclusiveness in design."	Risks to the **right to equality and non discrimination** (Article 21, EU Charter of Fundamental Rights). Risk of perpetuating and **amplifying existing societal bias** and **discriminations** against certain collectives or minority groups. Risk of applying a false computer neutrality to an algorithmic decision based on biased datasets. This dimension ensures that algorithmic decisions do not display an unjust or biased behavior with respect to sensitive factors such as gender, race or religion. AI technologies have deep reach and can transform political, economic and social institutions of the 21st century. Used by and serving the interests of the powerful, whether it is the state or a corporate actor, artificial intelligence's design, development and deployment (AI-DDD) reinforces power structures and can enable oppression of the vulnerable rather than their protection and empowerment. (Fukuda-Parr & Gibbons, 2021). There is also a risk of lack of information sharing that leads to harm (inability to audit for bias in machine learning) (Benthall & Haynes, 2019).	The models used to analyse, interpret and communicate learning analytics to stakeholders (support staff, advisers, faculties, students) should be sound, free from algorithmic bias; transparent where possible and clearly understood by the end users (Slade & Tait, 2019).	Early warning systems powered by AI will typically profile students and identify who is at risk of dropping out. If their effectiveness in identifying the right students is too limited, even if they do no more harm than the lack of a system, they are not fully trustworthy and need improvement through further research and development. Another possibility is that they are accurate but misused. Identifying who is at risk of dropping out matters only if a good (human) intervention to support the students and address that risk is implemented. (Vincent-Lancrin and Van der Vlies, 2020).	Datasets that represent the field of possible users in training, testing, and validation sets. Including minority voices in model training. Ways to mitigate bias in training by educating developers on possible bias in their models. Using Universial Deisgn for Learning to allow all kinds of users to interact with the system. Stakeholders are consulted on the design of the technology.	Use participatory tools that could elicit information about how different groups experience discrimination and use those insights to explore what fairness and diversity would look like in the context of the AI initiative. Is your definition of fairness commonly used and implemented in any phase of the process of setting up the AI system? Did you consider other definitions of fairness before choosing this one? Did you consult with the impacted communities about the correct definition of fairness, i.e. representatives of persons with disabilities?
6. Societal and Environmental Well-being	"environmental responsibility"; "Impact on Work and Skills"; "Impact on Society at large or Democracy."	Ubiquitous exposure to social AI systems in all areas of our lives (be it in education, work, care or entertainment) may **alter our conception of social agency**, or **negatively impact our social relationships** and attachment. While AI systems can be used to enhance social skills, they can equally contribute to their deterioration. This could equally affect peoples' physical and mental well-being. (EC High-Level Expert Group on AI, 2020). There is also a risk of lack of information sharing that leads to harm (inability to identify harmful patterns that could be prevented through service interventions) (Doll, 1974; Parkin & Paul 2011).	Governments must work with stakeholders to shape AI in education to help prepare for the transformation of the world of work and society. (Vincent-Lancrin and Van der Vlies, 2020).	There may be some educational information related to progress and social connection that may lead a teacher or other education professional to recommend health or social service interventions. It may be necessary to use and share data for these purposes.	Developer considers technology's environmental or societal context, such as energy use and carbon emissions, and its impact on the humans who will use the technology. Technology that does not have negative impacts on democracy (such as by amplifying fake news).	Use participatory tools that could elicit information about under what circumstances education data and AI outputs could be used for secondary purposes to support meaningful service interventions that could improve wellbeing. Citizen juries (Parkin & Paul, 2011), embodied participatory methods (e.g., walkthrough or media go-along) (Malinverni et al., 2019; Light et al., 2018; Jørgensen, 2016)
7. Accountability	"verifiability and replicability," "impact assessments," "evaluation and auditing requirements," "creation of a monitoring body," "ability to appeal," "remedy for automated decision," "liability and legal responsibility," and "accountability per se."	Risks related to **interpretability** and **explainability**. The first refers to a measure of the white-boxiness of a model. The second seeks the verbalization of algorithmic decisions at different levels of abstraction, corresponding to the different knowledge and needs of stakeholders, regulators and end-users. It accounts for the risks of ensuring that algorithmic decisions can be contested and reasoned upon. (EU Charter of Fundamental Rights, Article 47, **Right to an effective remedy** and GDPR, Art. 13-15, **Rright to explanation**). The the idea of explainability often transcends the ML models themselves to include not only the technical but also the human dimension (Unceta, Nin and Pujol, 2020)	Explainability (sometimes called interpretability) that goes beyond satisfying students' desire to understand the application and legal requirements to provide explanations (GDPR, Art. 13-15, Rright to explanation). Explainability helps designers enhance correctness, identify improvements in training data, account for changing realities, support students in taking control, and increase user acceptance. (Weld and Bansal, 2019).	A student may wish to understand why they are receiving a certain rating on their progress or a particular support. They may wish to remedy a perceived data error.	Regular assessment of the tools built. Outside parties that can evaluate the tool for biases and effectiveness. Programmers and designers who understand the legal obligations of AI. Reliable human-centered AI systems are produced by applying sound technical practices to software engineering teams. These technical practices clarify human responsibility, such as audit trails for accurate records of who did what and when, and histories of who conducted design, coding, testing, and revisions. (Shneiderman, 2020).	Use participatory tools that could elicit information about how stakeholders would like to see the lines of accountability for the operation and recommendations of the AI tool. Contextualizing with scenarios (Scenario-based approach), sentiment analysis (Ingrams, 2020).

References

Fjeld, Jessica and Achten, Nele and Hilligoss, Hannah and Nagy, Adam and Srikumar, Madhulika, Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI (2020). Berkman Klein Center Research Publication No. 2020-1, Available at SSRN: https://ssrn.com/abstract=3518482 or http://dx.doi.org/10.2139/ssrn.3518482

European Commission. Directorate General for Communications Networks, Content and Technology. (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment: Publications Office. https://doi.org/10.2759/002360

Vincent-Lancrin, S. and R. van der Vlies (2020), "Trustworthy artificial intelligence (AI) in education: Promises and challenges", OECD Education Working Papers, No. 218, OECD Publishing, Paris, https://doi.org/10.1787/a6c90fa9-en.

Slade, Sharon & Tait, Alan. (2019). Global guidelines: Ethics in Learning Analytics.

Fengchun Miao, Wayne Holmes, Ronghuai Huang, Hui Zhang (2021). AI and Education: Guidance for Policy-Makers, UNESCO, ISBN 978-92-3-100447-6.

Bright, J., Ganesh, B., Seidelin, C. & Vogl, T. (2019) Data Science for Local Government. Oxford Internet Institute, University of Oxford.

Data Futures Partnership. (2017). A Path to Social Licence: Guidelines for Trusted Data Use. Data Futures Partnership. https://toi-aria.s3.amazonaws.com/documents/Summary-Guidelines.pdf

Centre of Excellence for Information Sharing. (2018). The General Data Protection Regulation—An opportunity for change. Centre of Excellence for Information Sharing. http://informationsharing.org.uk/gdpr/

Kenny, S. (2008, May 1). An Introduction to Privacy Enhancing Technologies. The International Association of Privacy Professionals (IAPP). https://iapp.org/news/a/2008-05-introduction-to-privacy-enhancing-technologies/

Doll, R. (1974). Public Benefit and Personal Privacy: The Problems of Medical Investigation in the Community. Proceedings of the Royal Society of Medicine, Symposium on Constraints on the Advance of Medicine, 67, 1281–85.

Prinsloo, P., & Slade, S. (2017). Ethics and learning analytics: Charting the (un) charted. SoLAR.

Ifenthaler, D., & Schumacher, C. (2016). Student perceptions of privacy principles for learning analytics. Educational Technology Research & Development, 64(5), 923–938. https://doi-org.libproxy.tuni.fi/10.1007/s11423-016-9477-y

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. British Journal of Educational Technology,. doi:10.1111/bjet.12152.

Unceta I., Nin J., Pujol O. (2020) Risk mitigation in algorithmic accountability: The role of machine learning copies. PLOS ONE 15(11): e0241286. https://doi.org/10.1371/journal.pone.0241286

Prinsloo, P., & Slade, S. (2014). Student privacy and institutional accountability in an age of surveillance. 10.1007/978-94-6209-794-0_12.

Ernst, A., Biß, K. H., Shamon, H., Schumann, D., & Heinrichs, H. U. (2018). Benefits and challenges of participatory methods in qualitative energy scenario development. Technological Forecasting and Social Change, 127, 245–257. https://doi.org/10.1016/j.techfore.2017.09.026

Shadowen, N., Lodato, T., & Loi, D. (2020). Participatory Governance in Smart Cities: Future Scenarios and Opportunities. In N. A. Streitz & S. Konomi (Eds.), LNCS Sublibrary: SL3 - Information Systems and Applications, incl. Internet/Web, and HCI: Vol. 12203. Distributed, ambient and pervasive interactions: 8th International Conference, DAPI 2020, held as part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, proceedings / Norbert Streitz, Shin'ichi Konomi (eds.) (Vol. 12203, pp. 443–463). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-50344-4_32

Seale, J. (2010). Doing student voice work in higher education: an exploration of the value of participatory methods. British Educational Research Journal, Vol. 36, No. 6, pp. 995–1015

Chambers, R. (2009). So that the poor count more: using participatory methods for impact evaluation. Journal of Development Effectiveness, 1(3), 243–246. https://doi.org/10.1080/19439340903137199

Buchanan, W., Imran, M., Pagliari, C., Pell, J., & Rimpiläinen, S. (2020). Use of Participatory Apps in Contact Tracing: Options and Implications for Public Health, Privacy and Trust. https://doi.org/10.17868/73197

Shneiderman B. (2020). Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. ACM Trans. Interact. Intell. Syst. 10, 4, Article 26 (December 2020), 31 pages. DOI:https://doi.org/10.1145/3419764

Weld, D. S., & Bansal, G. (2019). The challenge of crafting intelligible intelligence. Communications of the ACM, 62(6), 70–79. https://doi.org/10.1145/3282486

Giannouchos, T. V., Ferdinand, A. O., Ilangovan, G., Ragan, E., Nowell, W. B., Kum, H.-C., & Schmit, C. D. (2021). Identifying and prioritizing benefits and risks of using privacy-enhancing software through participatory design: A nominal group technique study with patients living with chronic conditions. Journal of the American Medical Informatics Association : JAMIA, 28(8), 1746–1755. https://doi.org/10.1093/jamia/ocab073

Harrington, C., Erete, S., & Piper, A. M. (2019). Deconstructing Community-Based Collaborative Design. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–25. https://doi.org/10.1145/3359318

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

Nevaranta, M., Lempinen, K., & Kaila, E. T. (2020). Students' Perceptions about Data Safety and Ethics in Learning Analytics. 1613-0073. Retrieved from https://helda.helsinki.fi/bitstream/handle/10138/325523/FP_2.pdf?sequence=1&isAllowed=y

Brayne, S. (2014). Surveillance and System Avoidance: Criminal Justice Contact and Institutional Attachment. American Sociological Review, 79(3), 367–391. https://doi.org/10.1177/0003122414530398

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. Government Information Quarterly, 101493. https://doi.org/10.1016/j.giq.2020.101493

Jakobsen, M., Petersen, N. B. G., & Laumann, T. V. (2018). Acceptance or Disapproval: Performance Information in the Eyes of Public Frontline Employees. Journal of Public Administration Research and Theory, 29(1), 101–117. https://doi.org/10.1093/jopart/muy035

Mumtaz, S. (2000). Factors affecting teachers' use of information and communications technology: A review of the literature. Journal of Information Technology for Teacher Education, 9(3), 319–342. https://doi.org/10.1080/14759390000200096

Benthall, S., & Haynes, B. D. (2019). Racial Categories in Machine Learning. Proceedings of the Conference on Fairness, Accountability, and Transparency, 289–298. https://doi.org/10.1145/3287560.3287575

Parkin, L., & Paul, C. (2011). Public good, personal privacy: A citizens' deliberation about using medical information for pharmacoepidemiological research. Journal of Epidemiology and Community Health, 65(2), 150–156. https://doi.org/10.1136/jech.2009.097436

Ingrams, A. (2020). A machine learning approach to open public comments for policymaking. Information Polity, 25(4), 433–448. https://doi.org/10.3233/IP-200256

Jørgensen, K. (2016). The media go-along: Researching mobilities with media at hand. MedieKultur: Journal of Media and Communication Research, 32(60). https://doi.org/10.7146/mediekultur.v32i60.22429

Light, B., Burgess, J., & Duguay, S. (2018). The walkthrough method: An approach to the study of apps. New Media & Society, 20(3), 881–900. https://doi.org/10.1177/1461444816675438

Malinverni, L., Schaper, M.-M., & Pares, N. (2019). Multimodal methodological approach for participatory design of Full-Body Interaction Learning Environments. Qualitative Research, 19(1), 71–89. https://doi.org/10.1177/1468794118773299

Viale Pereira, G., Cunha, M. A., Lampoltshammer, T. J., Parycek, P., & Testa, M. G. (2017). Increasing collaboration and participation in smart city governance: A cross-case analysis of smart city initiatives. Information Technology for Development, 23(3), 526–553. https://doi.org/10.1080/02681102.2017.1353946

Weyns, D. (2020). Towards a code of ethics for autonomous and self-adaptive systems. Proceedings of the IEEE/ACM 15th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, 163–165. https://doi.org/10.1145/3387939.3391567

Gotterbarn, D., Miller, K., & Rogerson, S. (1999). Computer society and ACM approve software engineering code of ethics. Computer, 32, 84–88. https://doi.org/10.1109/MC.1999.796142

Thomson, A. J., & Schmoldt, D. L. (2001). Ethics in computer software design and development. Computers and Electronics in Agriculture, 30(1–3), 85–102. https://doi.org/10.1016/S0168-1699(00)00158-7