# RESPONSIBLE AI FRAMEWORK FOR LEARNING ANALYTICS IN HIGHER EDUCATION INSTITUTIONS

Resources compilation

Document in progress - Proposal V0.1

September 2024

The Open University

KMi Knowledge Media Institute

UKRI Science and Technology Facilities Council

# RESPONSIBLE AI FRAMEWORK FOR LEARNING ANALYTICS IN HIGHER EDUCATION INSTITUTIONS

## Introduction

We introduce our Responsible AI (RAI) framework tailored to Learning Analytics (LA) in Higher Education (HE).

The primary aim of this framework is to provide higher education institutions with actionable guidance on how to incorporate responsible AI principles effectively into their LA initiatives.

Recognising that institutions are at various stages of their LA adoption, we have structured the framework to follow the stages of the software development lifecycle: Requirements and Data Collection, Design, Development, Testing, Release, and Monitoring. By aligning the framework with these stages, we address a key limitation of many existing resources, allowing HEIs to engage with the specific stage of development they are currently in.

This approach enables a more flexible, actionable pathway for integrating responsible AI principles.

## Content

While our ultimate goal is to provide both a list of actions HEIs can take to ensure their LA systems incorporate responsible AI principles and how to implement these actions, our literature review reveals a significant lack of real world examples of how HEIs have operationalised these principles—if they have done so at all. We acknowledge that this leaves our framework incomplete, particularly in offering specific, practical steps that have already been tested in the field. However, we see this as an opportunity for continued growth.

Our ambition is to refine this resource in collaboration with the wider academic and practitioner community, learning from best practices as they emerge. These slides are a summary of the proposed framework, with links to available documentation and real-world case studies from HEIs.

This evolving resource will allow the community to contribute relevant materials, such as code libraries, consent forms, and other practical examples, fostering a collaborative environment where institutions can learn from one another.

# FAIRNESS & BIAS

# Fairness & Bias

## Actions by stage

| | |
|---|---|
| **Requirements & data collection** | - Select a definition of fairness<br>- Identify relevant sensitive attributes (also consider intersectionality)<br>- Identify potential sources of bias (data bias, algorithmic bias, social bias)<br>- Identify potentially applicable bias identification and mitigation methods |
| **Design** | - Apply relevant pre-processing bias mitigation methods (e.g, manage data imbalances, errors or missing values)<br>- Analyse and select algorithms/parameters that could minimise biases for the given LA system<br>- Design system features that could minimise human biases (i.e., how humans interpret and act over the presented information) |
| **Development** | - Apply appropriate in-processing and post-processing bias mitigation methods<br>- Develop system features to minimise possible human biases (e.g, misinterpretations of the information) |
| **Testing** | - Select and apply appropriate evaluation metrics to ensure not just the accuracy of the LA system, but also its fairness<br>- When possible, test with different data to ensure robustness to data drifts.<br>- Test the algorithms behind the LA system, but also the real-life impact of the LA system as a whole (e.g, impact on student outcomes) |
| **Release & Monitoring** | - Continuously monitor the AL system for accuracy and fairness and the trade-off between them<br>- Continuously monitor advances in the AI fairness field (new metrics and methods) |

# Fairness & Bias

## Tools & examples

**AI FAIRNESS-> https://github.com/Trusted-AI/AIF360**

- AI Fairness 360 - Python: AIF360 is an extensible open-source library containing techniques developed by the research community to help detect and mitigate bias in machine learning models throughout the AI application lifecycle. This document will provide an overview of its features and conventions for users of the toolkit.

**Fairlearn (python) -> https://fairlearn.org/**

- **Fairlearn** is an open-source, community-driven project to help data scientists improve fairness of AI systems.

**WhatIfTool -> https://pair-code.github.io/what-if-tool/**

- Using WIT, you can test performance in hypothetical situations, analyze the importance of different data features, and visualize model behavior across multiple models and subsets of input data, and for different ML fairness metrics.

**Using AI Fairness360 & Fairlearn**

- Deng, W. H., Nagireddy, M., Lee, M. S. A., Singh, J., Wu, Z. S., Holstein, K. and Zhu, H. (2022) 'Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits', *2022 ACM Conference on Fairness, Accountability, and Transparency*. DOI: 10.1145/3531146.3533113 .

### Common accuracy metrics

- false positive rate (FPR), false negative rate (FNR), true positive rate (TPR), true negative rate (TNR), Positive predictive value (PPV)
  - Examples and definitions: Agarwal, S. and Mishra, S. (2021) Responsible AI: Implementing Ethical and Unbiased Algorithms. DOI: 10.1007/978-3-030-76860-7

### Fairness metrics

- Equal Opportunity, Predictive Equality, Equalized Odds, Predictive Parity, Demographic Parity, Average Odds Difference
  - Examples and definitions: Agarwal, S. and Mishra, S. (2021) Responsible AI: Implementing Ethical and Unbiased Algorithms. DOI: 10.1007/978-3-030-76860-7

### Bias audit toolkit

- **AEQUITAS.** University of Chicago's open source bias audit toolkit for machine learning developers. https://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/

### Software for detecting algorithmic discrimination

- Repositories: https://github.com/megantosh/fairness_measures_code/

### Ethics

- An ethics checklist for data scientists. https://deon.drivendata.org/

# Fairness & Bias

## Literature

**Unfairness mitigation:**

- Deho, O. B., Zhan, C., Li, J., Liu, J., Liu, L. and Duy Le, T. (2022) 'How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics?', *British Journal of Educational Technology*. DOI: 10.1111/bjet.13217.

**Fairness definitions**

- Verma, S. and Rubin, J. (2018) 'Fairness definitions explained', *Proceedings of the International Workshop on Software Fairness*. DOI: 10.1145/3194770.3194776 .
- Bayer, V., Hlosta, M. and Fernandez, M. (2021) 'Learning Analytics and Fairness: Do Existing Algorithms Serve Everyone Equally?', *Artificial Intelligence in Education*. DOI: 10.1007/978-3-030-78270-2_12.

**Sensitive attributes**

- Deho, O. B., Joksimovic, S., Li, J., Zhan, C., Liu, J. and Liu, L. (2023) 'Should Learning Analytics Models Include Sensitive Attributes? Explaining the Why', *IEEE Transactions on Learning Technologies*. DOI: 10.1109/TLT.2022.3226474 .

**Dataset drift on the fairness of LA**

- Deho, O. B., Liu, L., Li, J., Liu, J., Zhan, C. and Joksimovic, S. (2024) 'When the Past != The Future: Assessing the Impact of Dataset Drift on the Fairness of Learning Analytics Models', *IEEE Transactions on Learning Technologies*. DOI: 10.1109/TLT.2024.3351352 .

**Fairness metrics**

- Caton, S. and Haas, C. (2024) 'Fairness in Machine Learning: A Survey', *ACM Computing Surveys*. DOI: 10.1145/3616865 .

**Bias**

- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2022) 'A Survey on Bias and Fairness in Machine Learning'. Available at http://arxiv.org/abs/1908.09635 .
- Rabonato, R. T. and Berton, L. (2024) 'A systematic review of fairness in machine learning', *AI and Ethics*. DOI: 10.1007/s43681-024-00577-5 .

**NIST Standards**

- Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. https://doi.org/10.6028/NIST.SP.1270
-

# TRANSPARENCY

# Transparency

## Actions by stage

| | |
|---|---|
| **Requirements & data collection** | - Define who needs to be informed (staff, students, expert/novice technical audiences)<br>- Define what information needs to be delivered (which information is tracked by the LA system, and when, how is this data processed, which decisions are supported with such data, etc.)<br>- Define mechanisms for effectively delivering such information depending on the different roles (short videos, weakly emails, etc) |
| **Design** | - Design software features to be accompanied by relevant information (e.g., i icons)<br>- Design documentation of the LA system according to each relevant role (e.g., documentation for students vs. tutors, vs. software developers) Co-creation could help to better understand the information required for each role.<br>- Determine how and when to deliver information to the different roles so that it is clear and understandable, and it will effectively reach users)<br>- Design mechanisms so that users can request additional information and ask relevant questions |
| **Development** | - Implement communication campaigns to ensure the designed documentation is effectively distributed across relevant roles<br>- Incorporate explanations when appropriate within the LA system<br>- Implement mechanisms to enable users to ask further questions |
| **Testing** | - Assess whether users (students, tutors and other roles) are aware of all the different elements of the LA system (data, algorithm, etc.) and how these different elements affect them and other roles within and outside the HE Institution. Questionnaires, interviews, focus groups, and immediate feedback via the LA system are possible mechanisms to test how informed users are. |
| **Release & Monitoring** | - Continuously monitor whether users are well informed<br>- Consider feedback and further queries requested via the implemented mechanisms to modify the provided information<br>- Modify communication campaigns as needed to ensure more effective information mechanisms<br>- Adapt the LA system as needed to ensure different users can effectively process the provided information |

# Transparency

## Tools & examples

- **“Guidelines for designing and delivering a sufficiently interpretable AI system”**
(Lesli, 2019, p. 44) in ‘Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector’

- **The Data Cards Playbook**
https://sites.research.google/datacardsplaybook/
A toolkit for transparency in AI dataset documentation. The Data Card template captures 15 themes frequently look for when making decisions — many of which are not traditionally captured in technical dataset documentation.

- **AI FactSheets**
https://aifs360.res.ibm.com/
A project dedicated to support the collection relevant information (facts) about the creation and deployment of an AI model or service. Useful templates to capture this information: which facts are of interest and how should they be rendered.

## Literature

**Transparency and Learning Analytics**

- Veljanova, H., Barreiros, C., Gosch, N., Staudegger, E., Ebner, M. and Lindstaedt, S. (2023) ‘Operationalising Transparency as an Integral Value of Learning Analytics Systems – From Ethical and Data Protection to Technical Design Requirements’, Learning and Collaboration Technologies. DOI: 10.1007/978-3-031-34411-4_37

- Hakami, E. and Hernández-Leo, D. (2020) ‘How are Learning Analytics Considering the Societal Values of Fairness, Accountability, Transparency and Human Well-being? -- A Literature Review’.

- Scheffel, M., Tsai, Y.-S., Gašević, D. and Drachsler, H. (2019) ‘Policy Matters: Expert Recommendations for Learning Analytics Policy’, in Transforming Learning with Meaningful Technologies. DOI: 10.1007/978-3-030-29736-7_38.

# ACCOUNTABILITY

# Accountability

## Actions by stage

| | |
|---|---|
| **Requirements & data collection** | - Define roles & responsibilities<br>- Define consequences of misuse<br>- Define auditing approach |
| **Design** | - Define controls and guidelines within the LA system<br>- Set up manuals for users |
| **Development** | - Translate defined controls and guidelines into software features |
| **Testing** | - Test rights and responsibilities<br>- Make sure that implemented controls and guidelines are functioning correctly and providing expected outcomes |
| **Release & Monitoring** | - Perform periodic assessments<br>- Monitor changes in roles and responsibilities<br>- Monitor changes on consequences of misuse<br>- Ensure actions are taken as result or intended or unintended harms |

# Accountability

## Tools & examples

- **AI accountability examples and use cases**
  **https://aiethics.turing.ac.uk/modules/accountability/**
  Providing resources and training materials to help practitioners to establish an end-to-end accountability framework.

- **AI FactSheets**
  https://aifs360.res.ibm.com/
  A project dedicated to support the collection relevant information (facts) about the creation and deployment of an AI model or service. Useful templates to capture this information: which facts are of interest and how should they be rendered.

**Concepts**
https://alan-turing-institute.github.io/turing-commons/skills-tracks/aeg/chapter4/accountability/

- Accountability requirements:
  - answerability
  - auditability
- Types of accountability
  - Anticipatory accountability
  - Remedial Accountability

## Literature

**Operational criteria for accountability**

- Veljanova, H., Barreiros, C., Gosch, N., Staudegger, E., Ebner, M. and Lindstaedt, S. (2022) 'Towards Trustworthy Learning Analytics Applications: An Interdisciplinary Approach Using the Example of Learning Diaries', HCI International 2022 Posters, Communications in Computer and Information Science. DOI: 10.1007/978-3-031-06391-6_19.

**Audit measures**

- Reidenberg, J. R. and Schaub, F. (2018) 'Achieving big data privacy in education', Theory and Research in Education. DOI: 10.1177/1477878518805308 .
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020) 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing'.

**Designing projects and accountability**

- Patterson, C., York, E., Maxham, D., Molina, R. and Mabrey, P. (2023) 'Applying a Responsible Innovation Framework in Developing an Equitable Early Alert System: A Case Study', Journal of Learning Analytics. DOI: 10.18608/jla.2023.7795 .

**Accountability in AI concepts & definitions**

- Novelli, C., Taddeo, M. and Floridi, L. (2024) 'Accountability in artificial intelligence: what it is and how it works', AI & SOCIETY. DOI: 10.1007/s00146-023-01635-y.

# SAFETY

# Safety

## Actions by stage

| | |
|---|---|
| **Requirements & data collection** | - Define responsible design, development, test and deployment practices<br>- Perform tool safety risk assessment<br>- Define procedures to collect and maintain accurate data<br>- Secure staff training on responsible development practices |
| **Design** | - Define test scenarios to evaluate model and data accuracy<br>- Identify methods and metrics to evaluate model accuracy (e.g., error rate, accuracy) |
| **Development** | - Develop systems adhering to responsible development practices<br>- Implement monitoring and audit approaches to detect safety vulnerabilities (e.g.,data poisoning, misdirected reinforcement learning behaviour) |
| **Testing** | - Test predictions/outputs using testing scenarios<br>- Test for possible deviations from intended or expected functionality (e.g., test models data drifts) |
| **Release & Monitoring** | - Communicate responsible use of system, tools, predictions to end users<br>- Execute periodic safety assessment (e.g., monitor and detect data drifts, concept drifts)<br>- Evaluate and monitor impact of LA and predictions in students and staff |

# Safety

## Tools & examples

**Safety principle is strongly interlinked with other principles such as Fairness and Explainability**

Safety characteristics:

- Accuracy
- Reliability
- Robustness

Approaches:

- Accuracy and Performance Metrics

Risks posed to accuracy and reliability

- Concept Drift
- Brittleness

Risks posed to security and robustness

- Adversarial Attack: https://github.com/Trusted-AI/adversarial-robustness-toolbox
- Data Poisoning
- Misdirected Reinforcement Learning Behaviour

## Literature

- **Concrete Problems in AI Safety**
  Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mané, D. (2016) 'Concrete Problems in AI Safety'. Available at http://arxiv.org/abs/1606.06565 .

- **Considerations on safety for LA**
  Howell, J. A., Roberts, L. D., Seaman, K. and Gibson, D. C. (2018) 'Are We on Our Way to Becoming a "Helicopter University"? Academics' Views on Learning Analytics', Technology, Knowledge and Learning. DOI: 10.1007/s10758-017-9329-9.

**Recommendations to maintain up-to-date data**

- Prinsloo, P. and Slade, S. (2013) 'An evaluation of policy frameworks for addressing ethical considerations in learning analytics', Belgium.

# SECURITY

# Security

## Actions by stage

| | |
|---|---|
| **Requirements & data collection** | - Identify local and regional data security standards<br>- Gather security requirements and identify assets<br>- Perform a security risk assessment<br>- Create an information security policy |
| **Design** | - Use the security risk assessment to identify proper level of security measures against threats<br>- Incorporate appropriate security measures |
| **Development** | - Apply appropriate tools and measures to secure code, models and data<br>- Ensure secure configuration of tools |
| **Testing** | - Build routine security test for all elements of the system<br>- Test and review  security controls (mitigate or eliminate risks) |
| **Release & Monitoring** | - Monitor periodically compliance of security policies and improve them if necessary<br>- Apply security and monitoring response program for security threats and vulnerabilities |

# Security

## Tools & examples

- **AIRT360**
  **https://art360.res.ibm.com/**
  The open source Adversarial Robustness Toolbox provides tools that enable developers and researchers to evaluate and defend machine learning models and applications against the adversarial threats of evasion, poisoning, extraction, and inference.
  ART provides tools that enable developers and researchers to evaluate and defend machine learning models and applications against the adversarial threats of evasion, poisoning, extraction, and inference.

## Literature

### LA and GDPR

- Karunaratne, T. (2021) 'For Learning Analytics to Be Sustainable under GDPR—Consequences and Way Forward', Sustainability. DOI: 10.3390/su132011524 .

### Data Protection Frameworks

- Cormack, A. N. (2016) 'A Data Protection Framework for Learning Analytics', Journal of Learning Analytics. DOI: 10.18608/jla.2016.31.6 .

- Eleni, P. (2023) 'Towards a Secure and Privacy Compliant Framework for Educational Data Mining', Research Challenges in Information Science: Information Science and the Connected World, Lecture Notes in Business Information Processing. DOI: 10.1007/978-3-031-33080-3_35.

- Steiner, C. M., Kickmeier-Rust, M. D. and Albert, D. (2016) 'LEA in Private: A Privacy and Data Protection Framework for a Learning Analytics Toolbox', Journal of Learning Analytics. DOI: 10.18608/jla.2016.31.5.

# PRIVACY

# Privacy

## Actions by stage

| | |
|---|---|
| **Requirements & data collection** | - Perform privacy impact and risk assessment<br>- Determine nature of consent (opt-in/out)<br>- Define consent-seeking procedure<br>- Obtain consent for data collection<br>- Define what, how, and why data is being collected |
| **Design** | - Design methods for sharing knowledge on data literacy and protection (documents, workshops)<br>- Specify data access roles<br>- Specify anonymisation techniques |
| **Development** | - Implement data privacy techniques, considering use of privacy enhancing technologies<br>- Comply with transparency and communication regulations |
| **Testing** | - Ensure data access is appropriate for each stakeholder<br>- Ensure initially specified and later modified data preferences are appropriately propagated |
| **Release & Monitoring** | - Establish (independent) process for handling complaints regarding data access and use<br>- Continually monitor data access as roles and personnel change<br>- Deploy data literacy and protection training |

# Privacy

## Tools & examples

- **AI Privacy 360**
  **https://aip360.res.ibm.com/**
  The AI Privacy 360 Toolbox includes several tools to support the assessment of privacy risks of AI-based solutions, and to help them adhere to any relevant privacy requirements. Tradeoffs between privacy, accuracy, and performance can be explored at different stages in the ML lifecycle.

- **GDPR on Privacy**
  https://gdpr-info.eu/issues/privacy-by-design/

## Literature

**Policy**

- Scheffel, M., Tsai, Y.-S., Gašević, D. and Drachsler, H. (2019) 'Policy Matters: Expert Recommendations for Learning Analytics Policy', in Transforming Learning with Meaningful Technologies. DOI: 10.1007/978-3-030-29736-7_38.

**Data Privacy**

- Prinsloo, P., Slade, S. and Khalil, M., 2022. The answer is (not only) technological: Considering student data privacy in learning analytics.
- Eleni, P. (2023) 'Towards a Secure and Privacy Compliant Framework for Educational Data Mining', in Research Challenges in Information Science: Information Science and the Connected World. DOI: 10.1007/978-3-031-33080-3_35 .
- Patterson, C., York, E., Maxham, D., Molina, R. and Mabrey, P. (2023) 'Applying a Responsible Innovation Framework in Developing an Equitable Early Alert System: A Case Study', Journal of Learning Analytics. DOI: 10.18608/jla.2023.7795.

## Literature (cont.)

**Students' views on privacy**

- Gosch, N., Andrews, D., Barreiros, C., Leitner, P., Staudegger, E., Ebner, M. and Lindstaedt, S. (2021) 'Learning Analytics as a Service for Empowered Learners: From Data Subjects to Controllers', LAK21: 11th International Learning Analytics and Knowledge Conference. DOI: 10.1145/3448139.3448186 .
- Alzahrani, A. S., Tsai, Y.-S., Aljohani, N., Whitelock-wainwright, E. and Gasevic, D. (2023) 'Do teaching staff trust stakeholders and tools in learning analytics? A mixed methods study', Educational technology research and development. DOI: 10.1007/s11423-023-10229-w.
- Francis, M., Avoseh, M., Card, K., Newland, L. and Streff, K. (2023) 'Student Privacy and Learning Analytics: Investigating the Application of Privacy Within a Student Success Information System in Higher Education', Journal of Learning Analytics. DOI: 10.18608/jla.2023.7975 .

**Others**

- West, D., Huijser, H. and Heath, D., 2016. Putting an ethical lens on learning analytics. Educational Technology Research and Development.
- Tsai, Y.S., Whitelock-Wainwright, A. and Gašević, D., 2020, March. The privacy paradox and its implications for learning analytics. In Proceedings of the tenth international conference on learning analytics & knowledge .
- Alzahrani, A. S., Tsai, Y.-S., Iqbal, S., Marcos, P. M. M., Scheffel, M., Drachsler, H., Kloos, C. D., Aljohani, N. and Gasevic, D. (2023) 'Untangling connections between challenges in the adoption of learning analytics in higher education', Education and Information Technologies. DOI: 10.1007/s10639-022-11323-x .
- Clarke, R. (2018) 'Guidelines for the responsible application of data analytics', Computer Law & Security Review. DOI: 10.1016/j.clsr.2017.11.002 .

# EXPLAINABILITY

# Explainability

## Actions by stage

| | |
|---|---|
| **Requirements & data collection** | - Define what elements will need explanations (data, algorithm, outputs, …)<br>- Define the type of explanations needed considering different elements and users of the LA system |
| **Design** | - Define how explanations will be generated (e.g., by incorporating explainable AI tools, like LIME, by using explainable ML models like decision trees)<br>- Define how explanations will be presented to the user (e.g, via narratives, tables)<br>- Define how explanations will be deliver to the user (e.g, via interactive elements of the system, documentation)<br>- Define how the user will interact with such explanations (should they confirm they have read them, do they have an opportunity to challenge them) |
| **Development** | - Incorporate the selected explanation generation methods into the LA system<br>- Translate the output of the model explanation methods into different system, documentation and interface features to ensure users can see them and interact with them accordingly |
| **Testing** | - Test the correctness of the explanations<br>- Test the accessibility of the explanations (easy for users to find, and access) for different user groups<br>- Test the comprehensibility of explanations (easy for users to understand) for different user groups |
| **Release & Monitoring** | - Perform periodic assessments for correctness, accessibility and comprehensibility of explanations<br>- Consider changes based on user feedback and assessment results |

# Explainability

## Tools & examples

- **AIX360**
  https://aix360.res.ibm.com/
  This extensible open source toolkit can help you comprehend how machine learning models predict labels by various means throughout the AI application lifecycle. We invite you to use it and improve it.

- **LIME**
  https://github.com/marcotcr/lime
  Local Interpretable Model-agnostic Explanations (LIME).
  Lime is able to explain any black box classifier, with two or more classes. All we require is that the classifier implements a function that takes in raw text or a numpy array and outputs a probability for each class. Support for scikit-learn classifiers is built-in.

- **SHAP**
  https://shap.readthedocs.io/en/latest/index.html
  Shapley Additive Explanations is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions

## Literature

- **Explainable Artificial Intelligence (XAI): Concepts, taxonomies**
  Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. and Herrera, F. (2020) 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', Information Fusion. DOI: 10.1016/j.inffus.2019.12.012.

**Explainable AI**

- Lünich, M. and Keller, B. (2024) 'Explainable Artificial Intelligence for Academic Performance Prediction. An Experimental Study on the Impact of Accuracy and Simplicity of Decision Trees on Causability and Fairness Perceptions', The 2024 ACM Conference on Fairness, Accountability, and Transparency. DOI: 10.1145/3630106.3658953.

**Importance of XAI in LA context**

- Gunasekara, S. and Saarela, M., 2024. Explainability in Educational Data Mining and Learning Analytics: An Umbrella Review. In International conference on educational data mining. International Educational Data Mining Society.

**LIME and SHAP use**

- Li, M.J., Li, S.T., Yang, A.C., Huang, A.Y. and Yang, S.J., 2024. Trustworthy and Explainable AI for Learning Analytics. In LAK Workshops.

- Hlosta, M., Herodotou, C., Papathoma, T., Gillespie, A. and Bergamin, P. (2022) 'Predictive learning analytics in online education: A deeper understanding through explaining algorithmic errors', Computers and Education: Artificial Intelligence. DOI: 10.1016/j.caeai.2022.100108.
- Sachini Gunasekara and Mirka Saarela (2024) 'Explainability in Educational Data Mining and Learning Analytics: An Umbrella Review', International Educational Data Mining Society. DOI: 10.5281/ZENODO.12729987.

# REGULATIONS

# Regulations & Standards

## Compilation of AI related regulations & standards (work in progress)

**General Data Protection Regulation GDPR**
https://gdpr-info.eu/

**NIST**
**Artificial Intelligence Risk Management Framework (AI RMF 1.0)**
http://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

**STANDARDS FOR BIAS AUDIT**

- Foundational and terminological standards - ISO/IEC TR 24027
  What do we mean by "bias" and "fairness" in this context? What are we trying to measure?
- Process, management and governance standards - ISO/IEC 42001, ISO/IEC 23894
  What organisational and governance processes do you have in place to support responsible and fair innovation?
- Measurement and test methods - ISO/IEC TR 24027, ISO/IEC TS 12791
  What are the methods and metrics you're using to measure bias in your AI system?
- Product and performance requirements - ISO/IEC TR 24027, ISO/IEC 12791
  What is an acceptable output of my bias audit, in order for me to safely deploy this system? Is some bias acceptable?

**UK**
**Data ethics framework**
https://www.gov.uk/government/publications/data-ethics-framework

**Guide to Data Protection**
**ICO - Information Commissioner's Office**
https://ico.org.uk/for-organisations/

**Equality Act 2010**
https://www.legislation.gov.uk/ukpga/2010/15/contents

**Data Protection Act 2018**
https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

# Regulations & Standards

## Tools & examples

**Trustworthy and Ethical Assurance Platform**
**https://alan-turing-institute.github.io/AssurancePlatform/**
The Trustworthy and Ethical Assurance (TEA) Platform is an innovative, open-source tool designed to facilitate the process of creating, managing and sharing assurance cases for data-driven technologies, such as digital twins or AI.

**Ethics skills training**
https://alan-turing-institute.github.io/turing-commons/skills-tracks/

**AI Standards Hub**
https://aistandardshub.org/
The AI Standards Hub is a new UK initiative dedicated to the evolving and international field of standardisation for AI technologies. Find information on AI-related standards using the search and filtering capabilities below. This database currently covers nearly 300 relevant standards that are being developed or have been published by a range of prominent Standards Development Organisations. (Sept-2024)
*Note: The hub is an initiative in development, to date there are no links about LA, but we hope this changes in the coming years.*