STIAEM

Slow Tech Innovation

AI, Ethics and New Teaching Methods

Dataset Biases and Social Justice: Identifying Historical and Gender Distortions

Co-funded by the European Union

2025-1-IT02-KA220-SCH-000365213

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them... Project reference code: 2025-1-IT02-KA220-SCH-000365213

Università di Foggia

Studi Umanistici, Lettere, Beni Culturali, Scienze della Formazione

İSTANBUL - ESENLER

İLÇE MİLLİ EĞİTİM MÜDÜRLÜĞÜ

1474

STIAEM

Slow Tech Innovation AI, Ethics and New Teaching Methods

Co-funded by the European Union

2025-1-IT02-KA220-SCH-000365213

eduOpen

CC BY-NC-SA 4.0 International
This license requires that reusers give credit to the creator. It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, for noncommercial purposes only. If others modify or adapt the material, they must license the modified material under identical terms.

Università di Foggia
Studi Umanistici, Lettere, Beni Culturali, Scienze della Formazione

OER 3.3.1: Few-Shot prompting strategies: using pre-solved exemplar models for output stability
Reference Module: Module 3 - AI Ethics and Systemic Biases
Related Video-Pill: Video 3.3 - Fairness, Biases and Cognitive Effort
Authoring partner: Directorate of Secondary Education of Dodecanese - Greece
Topic / Thematic Focus: Few-shot prompting strategies, pre-solved exemplar models and output stability

Introduction and Learning Objectives

This resource accompanies video-pill 3.3 "Fairness, Biases and Cognitive Effort", together with OER 3.3.2, and addresses the theme of algorithmic justice. The idea to start from is simple and uncomfortable: an AI system is not neutral. It learns from the data we give it, and data carry with them the traces of the inequalities of the world in which they were produced.
The text connects to DigComp 3.0 (Cosgrove & Cachia, 2025), to the UNESCO (2021) perspective on rights and inequalities, and to the "AI pedagogy" pillar of the framework for teachers (Miao & Cukurova, 2024). Slow Tech and UDL orient the reading towards an equity designed for everyone, not for the average.
By the end of the reading, teachers will be able to:
identify the main types of bias along the data lifecycle (historical, representation, measurement, algorithmic);
explain, through the example of Gender Shades, why the average hides injustice and why disaggregation is a practice of justice;
recognise that algorithmic fairness entails value choices and translate them into criteria for the school.

1. Algorithmic neutrality does not exist

A system that learns from data reproduces the regularities present in the data, including unjust ones. Research on fairness defines the concept in negative terms, that is, as the absence of favouritism:
"fairness is the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making" (Mehrabi et al., 2021).
From this definition it follows that neutrality is not the starting state, but an objective to be pursued actively. A model can be technically accurate and, at the same time, systematically penalising for some groups. And the literature demonstrates that some fairness objectives are incompatible with each other: it is impossible to satisfy simultaneously all definitions of fairness, except in very particular cases. Equity, then, is not a neutral formula: it is a choice.

2. The bias cycle: a taxonomy

Mehrabi and colleagues propose a taxonomy useful for schools, which places biases along the process that goes from data to model. Historical bias concerns the world even before collection: already-existing inequalities that enter the data even with accurate sampling. Representation bias arises from the way one samples: some subpopulations are under-represented, and for them the model learns worse. Measurement bias depends on how one chooses and measures what one wants to observe: if you measure a concept with the wrong

variable, the result will be distorted. Algorithmic bias, finally, may not be present in the data and may be added by optimisation, which privileges overall accuracy at the expense of minority groups.

Recognising these levels means being able to intervene: there is no single remedy, but a series of intervention points. This is the meaning of correction techniques in pre-processing (transforming the data to remove discrimination), in-processing (constraining the objective during training) and post-processing (recalibrating the decision thresholds downstream). Schools do not implement them directly, but know them well enough to be able to ask informed questions of providers.

"Historical bias is the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection" (Mehrabi et al., 2021).

"Representation bias arises from how we sample from a population during data collection process" (Mehrabi et al., 2021).

"Pre-processing techniques try to transform the data so that the underlying discrimination is removed" (Mehrabi et al., 2021).

"In-processing techniques try to modify and change state-of-the-art learning algorithms in order to remove discrimination during the model training process" (Mehrabi et al., 2021).

"Post-processing is performed after training by accessing a holdout set which was not involved during the training of the model" (Mehrabi et al., 2021).

These statements offer schools a grid of questions to put to a provider: do the data reflect the world with its inequalities? Does the sample represent all relevant subpopulations? At which moment - before, during or after training - is intervention made to correct the distortion? Knowing that there are three moments of correction helps to assess the seriousness of a system, and not to settle for the generic claim "our model is fair".

3. Gender Shades: the average that hides injustice

The most effective demonstration of this mechanism is the Gender Shades study by Buolamwini and Gebru (2018), which analysed the main commercial gender classification systems on images of faces. The key result concerns the distribution of error:

"darker-skinned females are the most misclassified group (with error rates of up to 34.7%)" (Buolamwini & Gebru, 2018).

The figure acquires meaning in comparison with the most advantaged group: "The maximum error rate for lighter-skinned males is 0.8%". The difference is explained by looking at the training datasets, which are strongly imbalanced: "these datasets are overwhelmingly composed of lighter-skinned subjects (79.6% for IJB-A and 86.2% for Adience)". Dark-skinned people were so under-represented that the model had not learned to recognise them.

But the decisive passage is disaggregation. Looking at results separated by group, the authors show that the weight of error is concentrated on a single subpopulation: dark-skinned women, though making up 21.3% of the verification sample, produced between 61.0% and 72.4% of all errors. The overall average, apparently acceptable, hid a profound injustice. This is the point worth bringing into the classroom: it is not enough to know how much a system errs, one must know who it errs on.

"It is important to note that according to [83], it is impossible to satisfy some of the fairness constraints at once except in highly constrained special cases" (Mehrabi et al., 2021).

This result is uncomfortable but liberating. Uncomfortable, because it shows that there is no technical solution capable of satisfying at the same time all definitions of equity: calibration and class balancing, for example, are in tension with each other. Liberating, because it removes from schools the illusion that the issue is "solved" by the algorithm: equity remains a value choice, and as such belongs to public discussion and educational deliberation.

4. The methodological lesson: disaggregate and discuss

Gender Shades teaches a method even before a datum. First: always disaggregate results by group, and in an intersectional way, that is, by crossing the relevant variables (in the case of the study, gender and phenotype). Looking only at aggregates hides the most fragile groups. Second: be suspicious of the average, which is a synthesis and not a reality. Third: ask questions of justice before asking questions of efficiency, because a system can work well on average and be unjust for someone.

Then there is the question of values. Since not all definitions of equity can be satisfied together, every distributive choice requires debate. The school is the natural place to conduct it: not "the technical question", but "what kind of equity do we want for our community?". The ALTAI (European Commission, 2020) offers a self-assessment list to translate ethical principles into operational questions, useful also for those who do not design systems but adopt them.

On the practical plane, the European Union's ALTAI framework (European Commission, 2020) offers a self-assessment list that translates ethical principles - inclusion, non-discrimination, equity - into operational questions. Even those who do not design systems but adopt them can use its logic: asking, for each tool, whether it favours accessibility, whether it treats different groups equitably, whether a channel exists for reporting a discriminatory impact. For schools, algorithmic justice is not a chapter of computer science: it is a dimension of civic education.

5. Algorithmic justice in the classroom and at school

In the classroom, algorithmic justice becomes digital civic education: showing how biases arise, getting students to recognise them in concrete examples, training them to disaggregate data. An accessible exercise is to compare the performance of a tool on different groups within one's own school, or to analyse a public case (such as Gender Shades) asking which groups were missing from the data and what consequences follow.

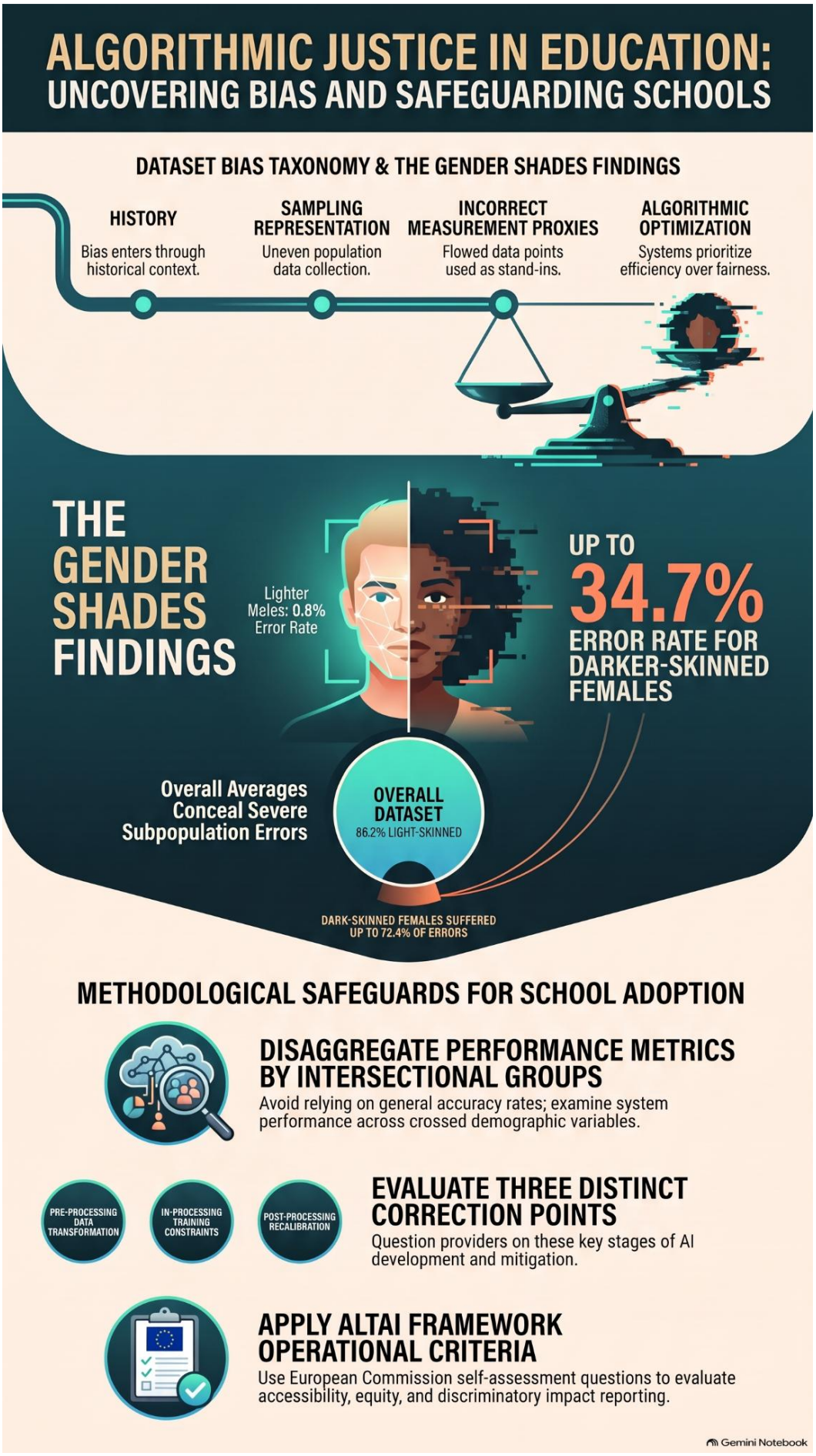
In the school as an organisation, algorithmic justice guides adoption choices. Before using a tool that assesses or guides, one must ask: on which data was it trained? how does it behave for our groups? what are the disaggregated metrics? These questions, if asked systematically, change the way tools enter the school system. The stakes are educational justice: ensuring that no group is penalised by the operation of a machine. The principle to keep in mind is that the average is not reality.

One way to make these concepts concrete is to start from one's own school's data. How many students belong to the groups that, in a case like Gender Shades, turned out to be penalised? If a system assessed or guided precisely those groups, how would we notice? Building scenarios of this kind, also in the form of role play or case analysis, develops in students - and in teachers - a habit of looking beneath the surface of aggregates.

The permanent availability of information online tends to shift attention from memory to search, with measurable cognitive effects (Sparrow et al., 2011).

Type of *bias*	Where it arises	Example	Safeguard
Historical	In the world, before collection	Social inequalities reflected in the data	Awareness and contextualisation
Representation	In sampling	Under-represented groups (79.6% / 86.2% light skin)	Balancing of *datasets*
Measurement	In the choice of variable	Measuring a concept with the wrong proxy	Validity of the constructs measured
Algorithmic	In optimisation	Average accuracy at the expense of minorities	Fairness constraints in training

Visual Scaffolding



Conclusions and Metacognitive Reflection Prompts

Algorithmic justice begins with recognising that no system is neutral. Biases have distinct places of origin - the world, sampling, measurement, optimisation - and require differentiated safeguards. Gender Shades shows that the way not to lose anyone is to disaggregate: to look at results by group, in an intersectional way, instead of stopping at the average. Since not all equities are compatible, choosing is an act of value, and the school is the place to discuss it.

To interrogate your own practice and your own tools:

- Which educational decision do you entrust (or would you entrust) to an automatic system, and which groups would risk being penalised?
- How would you bring into the classroom an exercise in data disaggregation accessible to your subject?
- Which questions of justice - beyond those of efficiency - would you put to a provider before adopting a tool?

Technical Metadata Sheet

Metadata Field	Specification / Parameter Value
Identifier	OER 3.3.1
Title	Dataset Biases and Social Justice: Identifying Historical and Gender Distortions
Creator	Directorate of Secondary Education of Dodecanese - Greece
Description	The resource examines algorithmic justice, starting from the idea that an AI system is never neutral because it learns from data that carry the traces of existing inequalities. It maps the main bias types across the data lifecycle and uses the Gender Shades case to show why averages hide injustice and why disaggregation is a practice of justice. The practical focus is on turning value choices about fairness into criteria for classrooms and schools.
Keywords	dataset bias; algorithmic fairness; social justice; representation bias; measurement bias; disaggregation; Gender Shades; digital equity
Language	EN (English)
Affiliation	Directorate of Secondary Education of Dodecanese - Greece
Contributor	STIAEM Transnational Scientific Board (peer review, Rhodes validation protocol)
Media Format	Text; Infographics
Genre	Training Manual; Educational Guide
URL	
Citation	Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 1-15. https://proceedings.mlr.press/v81/buolamwini18a.html Cosgrove, J., & Cachia, R. (2025). DigComp 3.0: European digital competence framework (Fifth edition). Publications Office of the European Union. https://data.europa.eu/doi/10.2760/0001149 European Commission. (2020). Assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115. https://doi.org/10.1145/3457607

	Miao, F., & Cukurova, M. (2024). AI competency framework for teachers. UNESCO. https://doi.org/10.54675/ZJTE2084
Educational Level	Secondary School
Associated Module / MOOC Video Lesson	Module 3 - AI Ethics and Systemic Biases / Video 3.3 - Fairness, Biases and Cognitive Effort
Learning Objectives	identify the main types of bias along the data lifecycle (historical, representation, measurement, algorithmic); explain, through the Gender Shades case, why the average hides injustice and why disaggregation is a practice of justice; recognise that algorithmic fairness entails value choices and translate them into criteria for the school.
Rights	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
Date Available	14/09/2026
Date Modified	

References and Webography

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 1-15. https://proceedings.mlr.press/v81/buolamwini18a.html

Cosgrove, J., & Cachia, R. (2025). DigComp 3.0: European digital competence framework (Fifth edition). Publications Office of the European Union. https://data.europa.eu/doi/10.2760/0001149

European Commission. (2020). Assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115. https://doi.org/10.1145/3457607

Miao, F., & Cukurova, M. (2024). AI competency framework for teachers. UNESCO. https://doi.org/10.54675/ZJTE2084

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. Science, 333(6043), 776-778. https://doi.org/10.1126/science.1207745