Educational AI and the Politics of Fairness: Structural Bias, Governance, and Student Empowerment

Zhengrui Cui *

School of Education, University of Birmingham, Birmingham, United Kingdom of Britain

* Corresponding Author Email: zxc488@student.bham.ac.uk

Abstract. With the widespread application of artificial intelligence systems in the field of education, algorithmic recommendations and student modelling are gradually replacing teachers and course designers in traditional teaching, becoming the core tools for the allocation of educational resources. However, the recommendation systems of existing educational platforms, while pursuing personalisation and improving efficiency through technological means, may exacerbate educational inequality. This paper analyses how educational platforms allocate resources through algorithms and examines recommendation systems to reveal the current state of student modelling and predictive pathways, particularly how algorithms use historical data and label features to shape students' learning trajectories. Subsequently, this paper explores the cognitive power structures underlying these predictive mechanisms and how algorithms profoundly influence students' self-perception and learning choices through data classification, ranking, and comparison. In summary, based on the above analysis, this paper proposes improvement suggestions from three aspects: policy, platform, and education, to enhance algorithm transparency and interpretability, design more inclusive learning paths, ensure educational equity, and protect students' cognitive autonomy..

Keywords: Educational artificial intelligence; algorithmic recommendation; student modelling; cognitive training; educational equity.

1. Introduction

In the context of platform-based learning becoming the norm, 'personalisation' in education is no longer merely about optimising the experience, but rather about opportunity allocation and path governance achieved through algorithms. Recommendation systems and student modelling transform learning into a measurable, predictable, and intervenable process: systems rank content, pace, and assessment under the guise of the 'optimal path'; students' abilities and potential are encoded as updatable states and guided accordingly [1-3]. This technological control is not neutral. As Metric Power reveals, measurement, classification, and ranking do not merely 'reflect' reality but also continuously shape learners' understanding of themselves and their future [4].

This paper focuses on two core questions: first, how do platform algorithms reproduce inequality along the 'recommendation-modelling-intervention' chain and erode students' cognitive sovereignty and negotiation space; Second, under the premise of not abandoning the benefits of automation, what governance conditions can transform student agencies into design constraints rather than post-hoc remedies. To this end, the one main body outlines the key mechanisms of educational platforms: recommendation logic, student modelling lineages (KT/BKT, IRT/MIRT, sequence depth models such as CAKT/CECAKT), and the cognitive power structure from 'prediction to intervention'. The other main body analyses how algorithmic allocation reproduces structural inequality through data loops and goal setting across four levels: individual, path, ideology, and group, and debunks the myth of 'algorithmic neutrality'. Chapter 4 proposes a three-tier governance framework: Policy/Institutional (transparency statements and independent reviews), Platform Design (explainability, counterfactual feedback, and longitudinal outcome tracking), and Educational Practice (elevating critical AI literacy from 'tool-based instruction' to systematic literacy in data, bias, and power), while emphasising that students' negotiable space serves as a hard constraint for algorithmic compliance and optimisation [5,6].

The aim of this paper is not to reject personalisation and automation, but to clarify their boundary conditions: algorithms can only truly serve educational justice--rather than redrawing the map of opportunity in the name of efficiency--when predictions can be questioned, paths can be rewritten, and consequences can be tracked.

2. Educational AI Mechanisms

2.1. Educational Recommendation Mechanisms and the Logics Behind Platform Automation

With the widespread application of artificial intelligence systems in the field of education, recommendation systems have become a key infrastructure for providing personalised learning resources, optimising instructional design, and enhancing learner engagement. These systems rely on various technical methods to match content with learners' needs, offering a powerful perspective for managing and experiencing educational content.

The most widely used method is Content-based recommendations, which compares various candidate items with items previously evaluated by users based on user profiles containing information about their tastes, preferences, and needs, and then recommends the most matching items; Collaborative recommendations, which predicts learners' preferences based on the past behaviour of similar users and attempts to predict the utility of a project for a specific user based on other users' previous evaluations of the project; and Hybrid Methods, which combine both techniques to enhance accuracy. Some recommendation systems adopt hybrid methods by combining Content-based recommendations and Collaborative recommendations, which helps avoid certain limitations of content-based and collaborative systems. Some platforms also utilise deep learning models to dynamically adjust recommendations based on students' real-time interactions and performance [7].

2.2. Student Modelling and Predictive Pathways

In contemporary educational platforms, there is an increasing reliance on data-driven educational systems. Students' learning paths are no longer entirely determined by personal experience or teacher judgment but are instead predicted and guided by a series of algorithmic models. The platform continuously collects learners' behavioural data to build models of their cognitive states, preferences, and potential needs, thereby adjusting content recommendations, task assignments, and even assessment methods.

The current mainstream student modelling methods can be broadly categorised into three types. The first type is the Knowledge Tracing model. In recent years, whether knowledge tracing can be effectively expanded to capture question difficulty and improve prediction accuracy has become an important issue to be addressed [1]. KT can be used to establish a student's knowledge model and determine when a student has mastered a particular skill. The reason for adopting the knowledge tracing method is that it serves as a form of cognitive diagnosis. For example, Baker's contextual guess and slip model is one such approach. Bayesian Knowledge Tracing (BKT) uses parameters to determine the probability of a student mastering a skill based on the temporal sequence of their correct and incorrect responses to a skill-related question, as well as its variant, the KT-IDEM model (Item Difficulty Effect Model) [1]. The second category involves specifying mathematical functions to describe the interaction between individuals and test items. This method standardises assessment data to provide a multidimensional characterisation of students' latent abilities, making it suitable for large-scale examinations, university admissions, and ability diagnosis applications [2]. Representative models include Item Response Theory (IRT) and Multidimensional Item Response Theory (MIRT). IRT focuses more on the age-related trends and underlying structure of student ability, helping systems predict future ability development. In recent years, researchers have also attempted to combine IRT with deep learning models to leverage their complementary roles in educational assessment and pathway planning. The third category of models typically combines sequence modelling (LSTM, Transformer) with convolutional network architectures to generate highly dynamic representations of students' knowledge states. This model posits that additional practice on the same knowledge concept enhances an individual's mastery of that concept. Representative models include the Convolution-Augmented Knowledge Tracing (CAKT) model and the Capsule-Enhanced

CAKT (CECAKT) model, which enable learning curve modelling. These models can learn both the overall knowledge state of students and the knowledge state of concepts in the next question [3].

2.3. From Prediction to Intervention: The Cognitive Power Structure in Learner Modelling

The aforementioned model is not only used for predicting student behaviour but also plays an invisible role in the planning and decision-making process of learning pathways. Students' 'future possibilities' are concretised into 'optimal pathways' through algorithmic reasoning, and the platform then provides feedback, recommendations, and interventions based on this. This means that the student modelling mechanism itself has quietly become an important tool for cognitive governance. It is therefore necessary to further explore the functional roles these modelling systems assume in practice, as well as their profound impact on students' cognitive pathways, goal setting, and educational decision-making. Although learner modelling systems are typically viewed as technical means to enhance personalisation, the predictive logic they rely on is not neutral but embedded with specific value orientations and behavioural assumptions.

These models often implicitly assume an 'ideal learning trajectory', treating certain learning methods as the standard, which may benefit some students but put others at a disadvantage, especially those with different starting points or diverse learning paths. Since they are trained using historical data, differences related to socioeconomic background, gender, or past academic performance are easily replicated or even amplified. More importantly, many platforms simplify learning objectives into efficiency and measurable outcomes, neglecting the exploratory and contextual differences that may exist in the learning process. As a result, systems that should provide students with diverse choices may inadvertently narrow their educational space, making it increasingly difficult for students to define and pursue learning paths based on their interests or potential.

This mechanism is closely related to Beer's Metric Power theory, which has been developed and spread across various social domains. Beer argues that data systems not only measure reality but also shape it through classification, ranking, and comparison, thereby influencing people's understanding of their own cognition and position. In an educational context, this means that learners' abilities, potential, and pace are gradually encoded as a set of 'optimal paths' within the system, and students gradually internalise these 'optimal' structures through continuous feedback and adjustment. This closed-loop mechanism of measurement-prediction-guidance may inadvertently weaken students' exploratory abilities and self-regulatory space. Metric Power highlights this phenomenon, where classification, ranking, and comparison generated through measurement gradually shape people's behaviour, choices, and self-understanding. In the educational domain, this power manifests as: students' cognitive paths are encoded as 'optimal trajectories,' and platforms use 'recommendations' as a guide for actual path regulation, ultimately potentially limiting students' active control over their own pace and interests. Therefore, student models and predictive path mechanisms not only enhance the intelligence level of the education system but also constitute a new form of governance over individual students. As technology is granted decision-making authority, issues related to students' cognitive sovereignty and educational equity require further scrutiny.

3. Structural Reproduction of Educational Inequity

3.1. Algorithmic Allocation and Structural Bias

Platforms often use students' past learning behaviour, label characteristics, and participation trajectories to make 'personalised' resource recommendations, adjusting the difficulty of the content they encounter, the frequency of feedback, and their learning pace. On the surface, this mechanism improves the efficiency of content distribution and individual suitability, but in reality, it may unknowingly replicate and reinforce existing inequalities. Previous studies have found that algorithms often use existing labels to categorise students and make recommendations [8]. Students with higher starting points and more active performance are more likely to be recommended advanced tasks and supplementary materials, while those with lower starting points or insufficient initial effort often get stuck in repetitive basic content. Such tiered recommendations may boost platform activity and retention rates in the short term, but in the long run, they may limit students' development potential and lead to the 'hierarchisation of information' in educational resources [9].

A more critical issue is that this allocation logic based on 'predictive ability' rapidly amplifies and solidifies students' initial minor differences into long-term learning trajectories, gradually narrowing their choices in education. Benjamin points out that algorithms are not merely knowledge transmitters [10]. They also embed value judgements into the path allocation process, thereby influencing students' understanding of their own abilities, future possibilities, and 'ideal goals'. Under such a mechanism, algorithms have effectively become a significant intervening force in the landscape of educational equity.

3.2. The Myth of Algorithmic Neutrality

Teachers increasingly rely on algorithmic judgments in practice but have a limited understanding of their underlying logic. Some teachers use the student profiles generated by the platform as a basis for adjusting teaching strategies yet overlook the data sample biases and model parameter limitations underlying algorithmic predictions. As Tsai et al. point out, the 'black-box nature' of educational AI systems is blurring the boundaries between professional judgement and technological power [11].

Therefore, so-called 'algorithmic neutrality' is actually a disguise that conceals the profound role algorithms play in shaping educational values and guiding student behaviour. If this misunderstanding is allowed to continue, it will easily lead to educational inequality. To avoid this risk, it is necessary to strengthen AI literacy education in teaching practices and improve the transparency and explainability of algorithms at the platform level, so that teachers and students can see the assumptions and biases behind the prediction logic and retain the space for questioning and correction.

3.3. The Differential Impact of Educational AI on Different Groups

Although educational AI is hoped to 'bridge the educational gap,' the actual deployment of technology and feedback reveals significant imbalances. The socio-cultural position of students not only affects their ability to access and use these systems smoothly, but also largely determines how they are 'defined' and 'recommended' by the systems.

Take the urban-rural gap as an example: the widespread adoption of digital platforms has indeed improved content accessibility in some rural and educationally disadvantaged areas, but these platforms often fail to account for local students' language habits, learning pace, and cultural background. Algorithms, by default, apply uniform ability labels and learning speed predictions to all students, and this 'one-size-fits-all' logic exacerbates existing mismatches. Some students with weaker foundations thus face excessive cognitive burdens during the learning process and may even lose the motivation to continue.

Gender differences are also worth noting. As Williamson and Eynon point out, recommendation systems often inadvertently replicate societal biases [6]. For example, in career path recommendations, female students are more likely to be guided toward humanities or care-related roles, while male students are more frequently directed toward science and engineering fields. Such 'mild' algorithmic interventions, while appearing to be personalised, actually reinforce existing structural stereotypes.

Therefore, if educational AI is to truly promote fairness, it must acknowledge the differing needs of various groups and establish a 'difference-sensitive' algorithm design philosophy. This means incorporating factors such as cultural context, learning styles, and resource accessibility into model construction and recommendation logic, thereby avoiding the so-called 'neutral technology' masking the 'non-neutral outcomes' it produces.

4. Governing Educational AI: Fairness, Accountability, and Student Agency

4.1. Designing Fairness: Principles and Pitfalls

In the process of large-scale introduction of artificial intelligence in the field of education, algorithmic 'fairness' is no longer just an abstract ethical issue, but a real-world problem directly related to whether students can obtain resources and opportunities. The so-called 'neutral processing' of algorithms is often misleading, as it often conceals the underlying logic of inequality. Currently, most educational platforms rely on predictive models to allocate learning resources and recommend learning paths, but these models are almost exclusively based on historical data. This means they cannot escape a fundamental paradox: when using past patterns to predict the future, they often end up replicating existing inequalities into future inequalities.

The 'target variable' in the education field--i.e., the object of algorithm optimisation--is inherently highly political [6]. For example, if the system aims to 'improve exam scores,' the algorithm is likely to over-concentrate resources on predictable 'high-performing students,' further marginalising groups with uncertain developmental trajectories. When optimising for 'student interest,' the system may underestimate cognitive potential due to insufficient modelling accuracy, leading to a mismatch between "ability" and 'interest'. This dilemma reveals that fair algorithm design in educational AI is not purely a technical issue but is deeply embedded in the social construction of 'what constitutes a good student' and 'what constitutes a successful path'. Additionally, Holstein et al. found in interviews with algorithm designers at technology companies that most designers could not clearly articulate the educational philosophy behind their fairness objectives [12]. They relied more on engineering feasibility than on educational ethical foundations, and this 'educational blind spot' easily solidifies biases in practice.

Educational data is inherently unstructured and highly heterogeneous, making it difficult to meet the balanced distribution required for high-quality machine learning. Training data often comes from past students' learning paths and behavioural records, meaning that historical inequalities (such as urban-rural gaps and cultural background constraints) are 'automatically accepted' during the modelling stage and further reproduced and amplified in the model outputs [13]. For example, if a particular group has consistently shown a 'low completion rate' in the past, the system is more likely to push them toward a 'simplified content' path, creating a cognitive feedback loop of 'low expectations--low recommendations--low achievements'. Suresh and Guttag argue that a fair system is not just about the model itself, but rather an entire process from goal setting, data collection, model construction, result presentation, to user feedback, where each link in the chain may produce bias. These biases are often masked by 'automated decision-making' mechanisms, leading to a broken feedback loop and making fairness impossible to implement.

4.2. Policy and Institutional Framework for Educational AI Governance

Against the backdrop of the continued expansion of educational artificial intelligence systems, students are increasingly becoming the objects of 'data-driven governance'. However, this governance process is often built on a technical structure that lacks student feedback mechanisms and participation rights. Holstein et al. point out that in mainstream teaching platforms, students typically cannot understand how the system evaluates their abilities, are unclear about why they are assigned specific tasks, and cannot refuse or adjust recommended learning paths. Such one-way data flow creates a 'black box' learning experience in practice, blurring the boundaries of system power and effectively constituting an implicit deprivation of students' cognitive sovereignty [12]. As Williamson and Eynon emphasise, when educational platforms deprive users of their bargaining space, algorithmic recommendations cease to be mere information services and instead become governance tools for constructing cognitive order [6]. As a result, educational AI gradually transforms into a 'micro-governance' mechanism that systematically reshapes students' understanding and behavioural patterns under the guise of personalisation, significantly eroding their learning agency and reflective space.

To break this invisible disciplinary structure, it is necessary to rebuild the feedback channels between students and AI systems. First, at the system interface level, buttons such as 'recommendation basis,' 'path explanation,' and 'content unsuitability' should be clearly set up to allow students to raise objections or provide preference feedback on AI recommendations. Empirical research by Hitron et al. indicates that students are more likely to trust AI systems that can 'explain their decision-making processes,' especially when it comes to long-term learning paths or exam tasks [12]. If the platform provides understandable feedback and options, user stickiness and learning autonomy are significantly enhanced. Secondly, a 'human-machine co-adjustment' mechanism should be introduced into the platform architecture to periodically assess the deviation between student feedback and actual performance, and adjust the recommendation algorithm accordingly, rather than sticking to the initial model.

It is necessary to systematically reconfigure the instructional design of Critical AI Literacy from three dimensions: course content, teaching objectives, and evaluation mechanisms. The AI4K12 framework proposed by Touretzky et al. provides a basic reference for educational institutions, but this framework still emphasises technical knowledge and tool usage [14]. Building on this, the following key dimensions should be further incorporated: Algorithmic bias and social justice: Guide students to analyse how AI amplifies social structural inequalities, such as gender discrimination and urban-rural disparities; The political nature of data: Through case analyses (career recommendations, educational pathway segmentation), help students identify issues of 'who is represented and who is excluded'; Platform responsibility and algorithmic governance: Discuss how platforms define students' capabilities, influence future choices, and whether students can request modifications to pathways or the deletion of records. Long and Magerko emphasise that AI education must transcend the 'technical learning' framework and shift toward 'cultivating awareness of power dynamics' [15]. Only when students truly understand AI's role in cognitive construction, value guidance, and social stratification can they question its logic while using the technology and actively participate in its optimisation and regulation. This educational objective is an indispensable cornerstone for advancing towards technological justice.

In summary, the construction of student feedback mechanisms and critical AI literacy courses should not be viewed as ancillary projects for 'user experience enhancement' or 'skill training,' but rather as core issues addressing the restructuring of educational power structures in the digital age. By embedding feedback mechanisms, it creates space for students to adjust their cognitive rhythms; through the design of critical AI literacy courses, it empowers students with the judgment and agency needed to navigate the uncertainties of future educational systems.

4.3. Rebuilding Feedback Channels: Student Agency in Platform Design

Algorithm systems widely deployed in current educational platforms have been deeply embedded in multiple stages, such as teaching recommendations, assessment feedback, and behaviour prediction. However, the algorithms of the vast majority of platforms remain in a 'black box' state, lacking effective external review and accountability mechanisms. In this context, establishing a systematic algorithmic regulatory mechanism is not only a necessary step to safeguard educational equity and student rights but also a core measure to address the absence of 'technological governance responsibility' [6].

The legal boundaries of algorithm governance should be established at the policy level. According to Holstein et al.'s industry survey, most educational platform developers lack a clear sense of ethical boundaries during design, instead relying on commercial objectives or technical feasibility to drive model logic [12]. This results in students being placed within algorithmic recommendation structures without their knowledge, with the collection and use of personal data remaining opaque. Therefore, it is necessary to establish institutional requirements such as 'algorithm transparency statements' and 'educational AI responsibility review mechanisms' to clarify the boundaries of educational data use and establish an independent review committee to conduct regular audits of platform algorithms.

Second, platforms should be encouraged to establish an 'explainability-first' mechanism. Hitron et al. point out that when students can obtain explanations of the recommendation logic, they are more likely to build trust in the system and maintain cognitive autonomy during use [12]. This mechanism not only enhances the system's user retention but also safeguards users' reflective space and feedback rights from a cognitive structural perspective. Therefore, the regulatory framework should include 'user explainability' as a compliance requirement for platform functionality, such as requiring the system to clearly state the reasons for recommendations, present decision-making paths, and provide 'counterfactual feedback' channels.

4.4. Embedding Critical AI Literacy in Education

Third, regulatory mechanisms must address the issue of 'dynamic evolution of algorithmic bias'. Unlike fields such as finance or healthcare, educational data systems exhibit distinct behavioural feedback loops (data loops): students' behaviour is influenced by recommendations, which in turn are based on prior behavioural feedback, thereby forming a dynamic process of path reinforcement. This mechanism can easily lead to the 'prediction-as-decision' effect, where biases embedded in early predictions may be amplified through multiple rounds of feedback, ultimately 'shaping' students' choices and self-perception [4]. Therefore, regulation should not only focus on static reviews of algorithmic structures but also establish a 'behavioural consequence tracking system' to monitor the distribution of algorithmic outcomes across different groups over the long term.

Finally, platform regulation is not merely a 'compliance' issue but also an 'educational ethics' issue. Foucault noted that the embedding of technology is always accompanied by the restructuring of disciplinary structures [16]. In the current context, algorithms, as a technological form of power, have quietly shaped students' learning rhythms, ways of thinking, and goal orientations. Therefore, regulation should not only focus on technical indicators but also involve an educational ethics committee in the platform design process, conducting a multi-dimensional assessment from the perspective of students' cognitive rights, autonomy, and exploration rights. This ethical embedding mechanism is an indispensable key component in achieving 'fair educational AI'.

In summary, the regulation of educational platform algorithms should not stop at formal compliance or external criticism but should establish a dynamic governance system centred on 'transparency--explanation--feedback--ethics'. This not only responds to the new challenges of educational equity in the digital age but also provides institutional safeguards

5. Conclusion

This paper takes algorithmic recommendation systems in educational platforms as its subject, systematically exploring how student modelling logic intervenes in the construction of cognitive structures and the allocation of educational opportunities through path recommendations, label classification, and behavioural prediction mechanisms. The study found that in current educational AI designs centred on the concept of 'optimisation,' student models not only serve a predictive function but also, under the influence of feedback mechanism deficiencies, path compression, and platform goal-orientedness, gradually solidify students' cognitive rhythms and developmental possibilities.

Additionally, the data logic and recommendation mechanisms relied upon by student modelling often overlook structural differences between social groups, replicating and amplifying educational inequality under seemingly neutral technical operations, thereby creating an 'algorithmic lock-in effect' where prediction becomes determination. This study further points out that addressing this issue requires not only technical optimisation but also multi-dimensional collaboration between institutional governance, platform design, and critical AI literacy education.

This paper adopts a perspective of 'cognitive regulation weight reallocation', attempting to break away from previous research paths that focused solely on algorithmic technology or ethical principles. It emphasises the critical role of the dynamic mechanism of 'feedback-explanation-participation' in educational AI systems for cognitive fairness. Future research could further explore algorithmic experience differences among various social groups in the actual use of educational platforms, driving the transformation of technological governance toward a multi-stakeholder negotiation model.

References

[1] Pardos Z A, Heffernan N T. KT-IDEM: Introducing item difficulty to the knowledge tracing model. User Modelling, Adaption and Personalization, 2011, 333-345.

[2] Reckase M D. Multidimensional item response theory. New York, NY, USA: Springer, 2009.

[3] Su H, Liu X, Yang S, Lu X. Deep knowledge tracing with learning curves. Frontiers in Psychology, 2023, 14: 1145999.

[4] Beer D. The social power of algorithms. in The Social Power of Algorithms, 2019, 1-13.

[5] Hitron T, Orlev Y, Wald I, Shamir A, Erel H, Zuckerman O. Can children understand machine learning concepts? in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, 1-11.

[6] Williamson B, Eynon R. Historical threads, missing links, and future directions in AI in education. Learning, Media and Technology, 2020, 45(3): 223-235.

[7] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734-749.

[8] Kizilcec R F, Saltarelli A J, Reich J, Cohen G L. Closing global achievement gaps in MOOCs. Science, 2017, 355(6322): 251-252.

[9] Veletsianos G, Houlden S. Radical flexibility and relationality as responses to education in times of crisis. Postdigital Science and Education, 2020, 2(3): 849-862.

[10] Benjamin R. Race after technology: Abolitionist tools for the new Jim code. Cambridge, UK: Polity Press, 2019.

[11] Tsai Y, Whitelock-Wainwright A, Gašević D. More than figures on your laptop: (Dis)trustful implementation of learning analytics. Journal of Learning Analytics, 2021, 8(3): 1-20.

[12] Holstein K, Vaughan J W, Daumé III H, Dudík M, Wallach H. Improving fairness in machine learning systems: What do industry practitioners need? in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, 1-16.

[13] Suresh H, Guttag J V. A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:1901.10002, 2019.

[14] Touretzky D, Gardner-McCune C, Martin F, Seehorn D. Envisioning AI for K-12: What should every child know about AI? in Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 9795-9799.

[15] Long D, Magerko B. What is AI literacy? Competencies and design considerations. in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, 1-16.

[16] Foucault M. Discipline and punish: The birth of the prison. New York, NY, USA: Vintage Books, 1977.