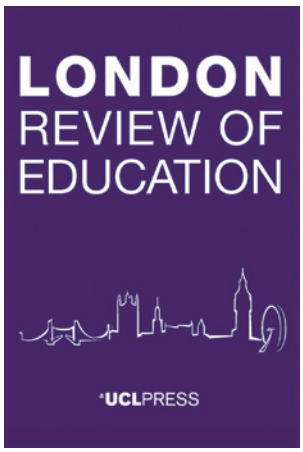

UCLPRESS

LONDON REVIEW OF EDUCATION

e-ISSN: 1474-8479

Journal homepage:
https://www.uclpress.co.uk/pages/london-review-of-education

Can artificial intelligence help predict a learner's needs? Lessons from predicting student satisfaction

Dimitris Parapadakis

How to cite this article
Parapadakis, D. (2020) 'Can artificial intelligence help predict a learner's needs? Lessons from predicting student satisfaction'. *London Review of Education*, 18 (2): 178–195. https://doi.org/10.14324/LRE.18.2.03

Submission date: 30 September 2019
Acceptance date: 20 March 2020
Publication date: 21 July 2020

Peer review
This article has been peer-reviewed through the journal's standard double-blind peer review, where both the reviewers and authors are anonymized during review.

Open access
London Review of Education is a peer-reviewed Open Access journal.

LONDON REVIEW OF EDUCATION

Parapadakis, D. (2020) 'Can artificial intelligence help predict a learner's needs? Lessons from predicting student satisfaction'. *London Review of Education*, 18 (2): 178–195. DOI https://doi.org/10.14324/LRE.18.2.03

Can artificial intelligence help predict a learner's needs? Lessons from predicting student satisfaction

Dimitris Parapadakis* – *University of Westminster, UK*

Abstract

The successes of using artificial intelligence (AI) in analysing large-scale data at a low cost make it an attractive tool for analysing student data to discover models that can inform decision makers in education. This article looks at the case of decision making from models of student satisfaction, using research on ten years (2008–17) of National Student Survey (NSS) results in UK higher education institutions. It reviews the issues involved in measuring student satisfaction, shows that useful patterns exist in the data and presents issues involved in the value within the data when they are examined without deeper understanding, contrasting the outputs of analysing the data manually, and with AI. The article discusses risks of using AI and shows why, when applied in areas of education that are not clear, understood and widely agreed, AI not only carries risks to a point that can eliminate cost savings but, irrespective of legal requirement, it cannot provide algorithmic accountability.

Keywords: artificial intelligence, algorithmic accountability, National Student Survey (NSS), higher education, decision making

Introduction

Decisions made in UK higher education, from course level to government policy and funding, increasingly make use of a wealth of data collected at all levels. According to UK government, a way forward that improves decision making is artificial intelligence (AI), which 'can be integrated into existing processes, improving them, scaling them, and reducing their costs, by making or suggesting more accurate decisions through better use of information' (Hall and Pesenti, 2017: 2). Machine learning, now used interchangeably with AI, is at the core of this approach. It aims to take part of the role of a researcher by analysing collected data to find unseen links and patterns of behaviour, providing greater understanding of why things are as they are. Machine learning has been around for many decades but has recently gained attention, with computer developments giving it the speed and scale to tackle real-life problems.

With machine learning, AI has changed the paradigm of software algorithms. Traditional software development will start with a human analysis of the problem, with discussion with domain experts, and proceed with the development of a series of algorithmic models of the experts' knowledge, so that the process can be replicated by a computer. The resulting algorithms are well understood by their developers, and they can be validated against the experts' knowledge. With machine learning, the analysis of the problem and the development of algorithms are all done by the computer; experts may not be involved. The problem may be too new and complex for

*Email: D.Parapadakis@westminster.ac.uk

anyone to have prior expertise in it, and, often, no developer or expert can have any understanding of how that algorithm works.

The wide interest in the use of AI in education is discussed in Zawacki-Richter *et al*. (2019), in areas ranging from admissions, retention, modelling and support of achievement, assessment and feedback, and also in supporting educators to get a better understanding of the learner's needs. That review rightly states that 'we should not strive for what is technically possible, but always ask ourselves what makes pedagogical sense' (ibid.: n.p.). Haenlein and Kaplan (2019: 10) address a major misunderstanding when they argue that AI's lack of prejudice 'does not mean that systems based on AI cannot be biased', making a good argument for regulation of AI. However, such valid concerns can mistakenly lead decision makers to believe that a good AI system can still give to education the benefits it has given to industry, as long as it follows the equivalent safeguards. AI is not a new tool in education: the International Artificial Intelligence in Education Society has supported research about AI in education since 1997. Yet Hinojo-Lucena *et al*. (2019) show that despite the breadth of AI applications for education, the research area is still in its early stages.

This article will show that although education has metrics that can be analysed to get insights about students, and AI can perform that analysis more efficiently, the lack of understanding of the technology coupled with a lack of agreement of what we really measure in many areas of education can lead AI to inaccurate, risky and counterproductive decision making, in a way that leaves no avenue through which to explore and provide accountability.

There are many education metrics that offer data that can be analysed by AI. The education metric used in this article is deliberately chosen as one that comes with controversy. Student satisfaction is a factor of growing importance influencing UK government policy and funding, and the strategic aims of the UK's higher education institutions (HEIs). The UK government's stated aim is that 'all students, from all backgrounds, receive a high quality academic experience' (OfS, 2018: 1). In parallel, there is ongoing concern about how UK higher education could provide better value for money (House of Commons Education Committee, 2018).

Since 2005, the National Student Survey (NSS) has provided a standardized annual mechanism to measure student satisfaction. The metrics it produces are used to inform decisions about perceived quality and about funding of education (Gunn, 2018). This makes the understanding, improvement and prediction of these metrics strategic goals for UK HEIs that wish to meet the needs of their learners and be aligned with government policy and funding decisions. Therefore, the socio-economic impact of the survey is considerable, steering decisions about education strategy, priorities of student support and senior management involvement in courses appearing to underperform. Interestingly, this is despite reported flaws in NSS approach and use (see Cheng and Marsh, 2010; Sabri, 2013). Furthermore, Frankham (2015) shows that decisions are not necessarily made on understanding of, or even trust in, the survey.

From the point of view of education research, AI can be seen as a hired researcher who can methodically analyse educational data of enormous scale at very high speed and at very low cost. With pressure to improve NSS results, AI becomes a candidate to replace an educational researcher to analyse the complex data faster and at a lower cost, and to provide decision makers with insights into student satisfaction. To evaluate if AI is a suitable replacement, this article checks what patterns a manual research analysis can detect in the NSS, whether these patterns can be used to help decision makers, and then if AI can also find patterns equally accurate and informative, and faster.

What is being measured

Using the National Student Survey

The NSS is an annual survey for all publicly funded HEIs in the UK. It collects student perceptions of factors affecting the students' own satisfaction. The basic list of satisfaction questions is standardized, which helps analysis and comparisons. (The list changed only once, in 2017.) NSS data are publicly available and typically analysed for a specific year (Bell and Brooks, 2018) or a short range of years (Langan *et al*., 2013).

The NSS is designed to inform prospective students, and so nudge HEIs to improve student experience. It is increasingly used to show each institution's position in the sector, relative to competitors, in terms of perceived quality. The importance of the NSS was further raised when the UK government included it in the Teaching Excellence Framework (TEF) (DfE, 2017), with the intention of linking government funding to the performance of institutions in metrics of student experience.

Despite courses varying their scores between years, and institutions appearing to be trapped in a cycle of celebrating their rising scores one year and expressing determination to improve after surprise drops the following year, the NSS does not appear to be an instrument of randomness, and there are underlying patterns in its data. For example, Yorke *et al*. (2014) identify distinct behaviour patterns among art and design students, while Agnew *et al*. (2016), analysing business school results in the 2014 NSS, identify three NSS questions as the most relevant to the measure of student satisfaction. Therefore, HEIs are right to seek information within their NSS results, whether to address the causes of student dissatisfaction or to address the causes of students' reported dissatisfaction at the time of the survey.

Understanding the National Student Survey

Despite the wealth of educational experience among HEI decision makers (individually and collectively), analysing data that contain patterns of behaviour and assessing the value of those patterns are not the same thing. There are many methodologies to perform an analysis, but assessing the value of the results requires a deeper understanding of what is being measured and the rough outline of what the underlying models should look like. Bell and Brooks (2018) show that NSS results will differ by subject, with humanities scoring higher than media studies and engineering, and argues that whereas some HEIs appear to have prioritized the improvement of assessment and feedback scores to positively impact on their overall satisfaction scores, the influence of this would be very small, especially compared with areas such as teaching quality and course organization. Langan *et al*. (2013) also confirm that predictors of behaviour exist in NSS data, and differ between subjects, drawing attention to institutional errors in comparing a subject area's performance to the institutional average. Similarly, Fielding *et al*. (2010) argue that the patterns of behaviour differ by subject sufficiently to reduce any usefulness of comparing results between subjects.

The decisions of HEIs are consciously influenced by several factors – including cost, strategic aims, competition and appetite for change – and we need to appreciate that student choices on a survey can also hide a very multidimensional model. Yet Buckley (2012: 50) reports that 'Some institutions benchmark their departments and courses internally, against each other and the institutional average', suggesting a high-level inability to understand and engage with the nature and complexity of the data model. Despite the standardization of the NSS data and the wealth of guidelines to be

found in the research, HEI decision makers will still analyse the NSS in a way that leads to strategic directions that are wrong in focus, and at times wrong overall.

A further issue with the NSS is capturing the quality measure of 'high student experience' through students' reported perception of their satisfaction. In a revealing experiment, Boehler *et al*. (2006: 748) argue that 'The finding that compliments produce greater satisfaction than did the feedback suggests that studies of interventions designed to improve feedback should include outcome measures other than measures of student satisfaction'. This is well known in the literature, with 'Dr Fox's lecture' (Naftulin *et al*., 1973) showing that learners can be 'seduced' by the style of teaching and can overlook the quality of their learning.

Further to the question of what the NSS measures and how it is interpreted, the survey organizers provided students with a diverse list of examples of 'inappropriate influence' (Ipsos MORI, 2019) that HEIs or individual staff may try to exert over students to change their scores inappropriately, providing a contact email address to report these. This highlights, at the least, the temptation for HEIs to compromise the quality of the data with which they are evaluated.

This uncertainty should be contrasted with typical applications where AI has shown impressive successes, such as predicting customer behaviour (Vafeiadis *et al*., 2015), text classification (Khan *et al*., 2010), spam detection (McCord and Chuah, 2011), face recognition (Sun *et al*., 2014) and even fruit classification (Zawbaa *et al*., 2014). In all these examples, the computer is typically looking at well-understood data, either to find and point out similarities or to discover how to differentiate between correct and incorrect examples of something. AI can study lots of oranges and apples and learn how to tell them apart – and so can human beings: the problem domain is understood well. These successes do not mean that AI can perform as well in areas where human beings have problems understanding underlying models.

Methodology and samples used

Data used

The NSS produces quantitative data, publicly available in various formats and levels of aggregation. To increase consistency between years, full-time student data have been selected for each of the years for each subject area in each HEI, and the UK Provider Reference Number (UKPRN) has been used to identify an HEI consistently over the years (even when institutions change names or merge).

Subject areas are presented in the data in three levels of detail, and the more specific subject Level 3 aggregation was used for all years. Although the coding of the subject areas has been changing recently, all subject areas were used with their original coding to help comparison. Regarding the population surveyed in each year, students are expected to respond to the survey in their final year, but there are outliers where students take longer to finish or take a year out. There is no reason why this would affect the results of one NSS question over another, or across years, to a degree that would bias the subject-based analysis. Further to this, should the survey responses in a particular course fall below a threshold, the data combine this with the previous year's, which, in effect, mean that some students' responses will influence two years' scores. The response rate itself varies across years and institutions but, across the data, does not favour HEIs or subject areas or years.

The exception to this was 2017, when a student boycott of the survey impacted more on some areas than others. In 2017 the questions also changed, with some being removed, some added and some reworded, and with their numbering changing. To

address this, the 2017 numbers were mapped to the 2008–16 numbering, where the questions were the same or almost the same, and given new codes where different. The size of the 2017 data within the decade is not large enough to impact on the analysis, but correlations referred to in this article as holding over the decade were looked at independently for each year, in effect making 2017 an independent validation check.

Factors at institution and subject level will affect the results in specific years. These include policy change, institutional restructuring, significant changes in student–staff ratio and changes in course leaders. An analysis of institutions by volatility was done in parallel to confirm that these changes, in themselves, do not bias the overall analysis (unless the subject examined was taught at very few institutions).

Sampling choices

Where institutions were grouped by size to differentiate between small and large providers, there were many ways the HEIs could be split. Basing the split on overall size of the institution would be less helpful than the overall size of the subject area, which students are more likely to experience through class sizes. Similarly, basing the split on the size in 2008 or 2017 would not reflect the changes in recruitment (and existence) of the subject area in different HEIs. The threshold used to split was the median institution size cumulatively over the decade (using the population size, rather than the size of responses). For the split by institution type, there were again many ways to select subgroups of HEIs; Russell Group and Million+ were chosen. The larger M25 grouping was not picked because its breadth of HEIs did not make it that different from the total set of HEIs.

The subject area selected for the experiment was '035 Computer Science'. Aside from being extensively researched by Parapadakis and Baldwin (2018), it was selected because it had a high enough population, spread widely geographically, across sizes and types of HEIs and the background of students has not been shown to differ in a way that influences student expectations. For example, the subject domain is the same in the UK and globally, and, unlike many other subjects, it is very unlikely to have students who are new to the subject mixed with students from families with generations of expertise in the subject.

Method and software used

The two analyses contrasted were correlations (coupled with informed interpretation of the data) and AI. The numbers examined were the satisfaction percentages for each set of subject, institution and year. This is always calculated in the NSS as the percentage of students who answered 5 (very satisfied) or 4 (satisfied). The balance between 5 and 4 has been found in parallel to contain interesting patterns of information regarding the NSS, but this was not included in this study. The correlations were made between the different questions and the question of overall satisfaction (Q22 until 2017, and Q27 in 2017, mapped to Q22 for this study), and software was written to produce heat-maps that a HEI decision maker can interpret to identify likely influencers of overall student satisfaction.

For the AI, the same data were used with two different machine learning algorithms (using the RapidMiner software). First, a tree learner was used to model and predict student satisfaction; this can produce a decision tree of the model discovered, which the HEI decision maker can read top-down to interpret and use. Complementing that, a Bayesian learner was used on the same data. This analyses probabilities of different data to predict student behaviour, but, although the success of these predictions can be estimated, the user will not be able to see on what basis these predictions are done.

To differentiate between high and low satisfaction, all scores above the median provider score for that year were considered 'satisfied' and all below were considered 'unsatisfied'. This choice could be criticized as too arbitrary, as decision makers seeing a drop in NSS scores from 3rd to 35th place are unlikely to ignore this simply because scores are still in the top half of the providers. However, within the context of this experiment, it provides a clear distinction that relates consistently to each year's scores, and AI can analyse patterns to predict it.

Outcomes of the experiments

Manually examined correlations

In the first analysis, a series of heat-maps are used to illustrate the results. Each heat-map displays the years on the *y*-axis and the NSS questions on the *x*-axis. The questions have been expanded to include the three categories of the Teaching Excellence Framework (TEF) – teaching, assessment and support – and the new questions introduced since 2017, each starting with an underscore to differentiate it from the previous questions with the same number. Each cell represents how closely that question correlates with the overall satisfaction question (Q22) on that specific year, redder shades signifying a higher correlation and bluer shades signifying a lower correlation (deep blue signifying no value). The highest correlation of Q22 is with itself, which shows as a consistent vertical red bar across the years. The categories of the questions are listed in Table 1.

Table 1: Categories of NSS questions

NSS questions	Category
Q01–Q04, _Q03	Teaching on my course (grouped as T1-T)
Q05–Q09	Assessment and feedback (grouped as T2-A)
Q10–Q12	Academic support (grouped as T3-S)
Q13–Q15	Organization and management
Q16–Q18	Learning resources
Q19–Q21	Personal development
_Q05–_Q07	Learning opportunities
_Q21–_Q22	Learning community
_Q23–_Q26	Student voice

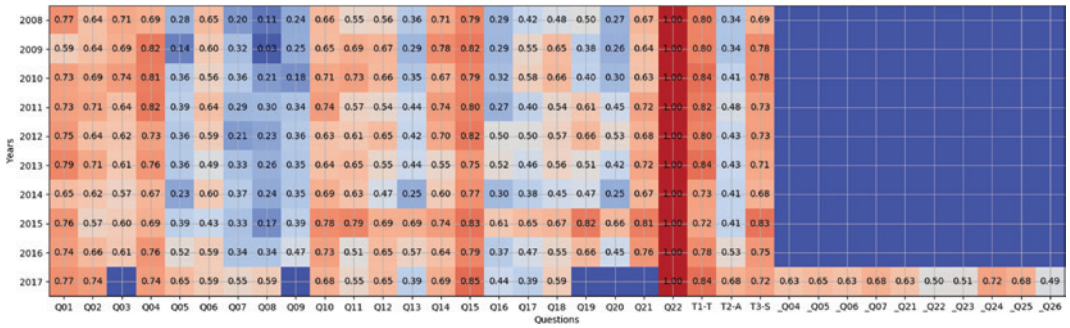

Figure 1: Heat-map of correlations of satisfaction of NSS questions with Q22, in computer science, per year, across all providers

The analysis of all computer science results of all institutions over all years (see Figure 1) shows strong correlations on teaching (T1-T, quite red across years) and support

questions (T3-S, also quite red), and a weak correlation of assessment and feedback (T2-A, mostly light blue), with the exception of Q06 (more red years than blue) about perceived fairness of assessment. Although Q04 (about the course being stimulating) correlates strongly for a few years, the consistent high performer is Q15 (about whether students believe their course is organized well and running smoothly): the Q15 column is showing consistently redder than other columns of questions. Therefore, although the heat-map shows that correlations can change between the years, it also shows consistent patterns that can be used to inform decision makers.

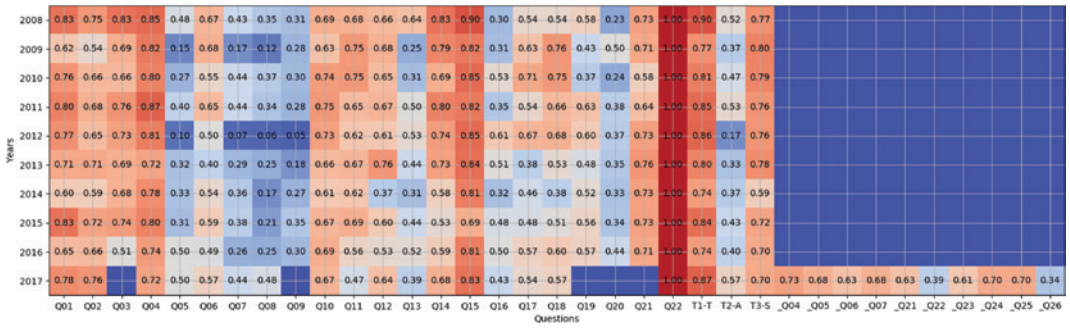

Figure 2: Heat-map of correlations of satisfaction of NSS questions with Q22, in computer science, per year, across the top half of providers by size

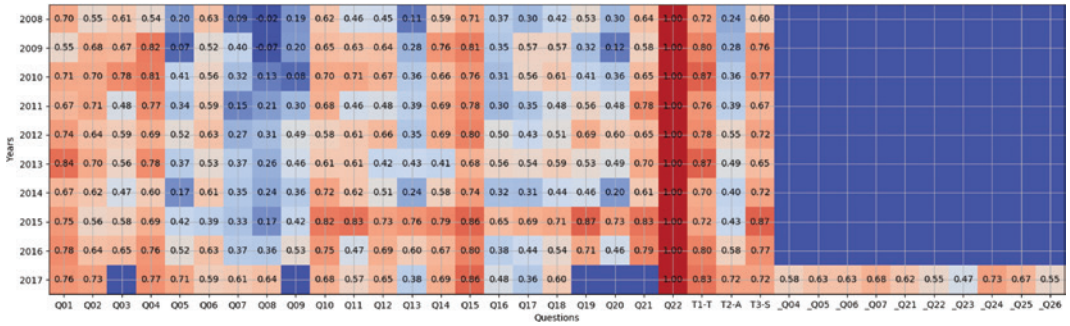

Figure 3: Heat-map of correlations of satisfaction of NSS questions with Q22, in computer science, per year, across the bottom half of providers by size

Looking at the same form of analysis but separating larger providers ('top half', see Figure 2) of computer science from smaller providers ('bottom half', see Figure 3), shows that correlations are generally weaker for the bottom half, but it is again evident that Q15 is strongly related to overall satisfaction, with Q04 a close second candidate.

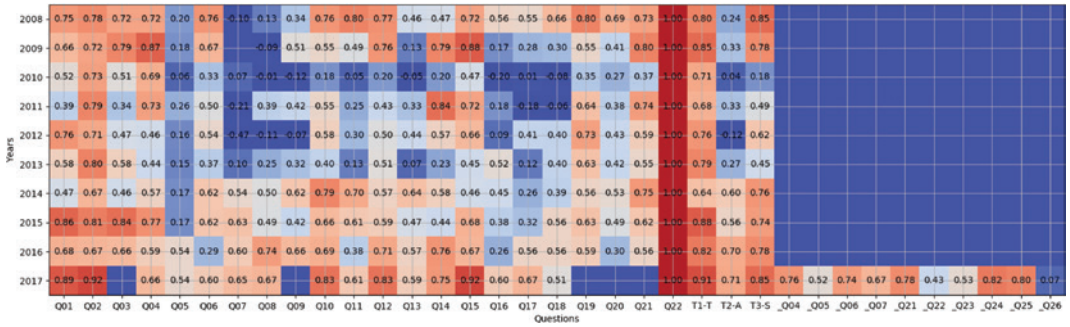

Figure 4: Heat-map of correlations of satisfaction of NSS questions with Q22, in computer science, per year, for Russell Group providers

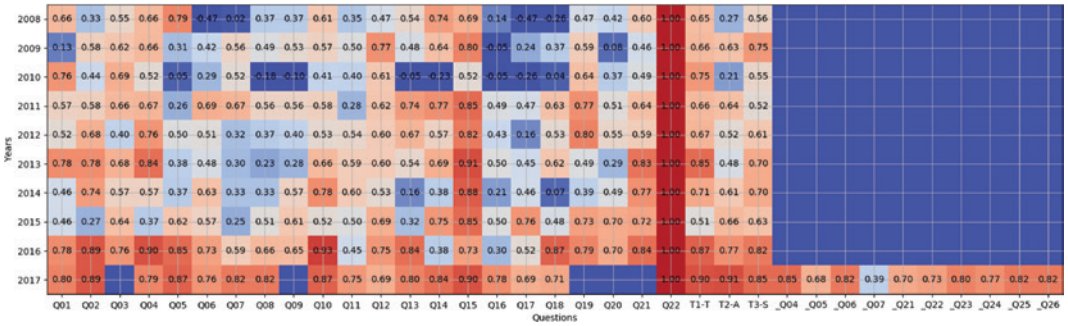

Figure 5: Heat-map of correlations of satisfaction of NSS questions with Q22, in computer science, per year, for Million+ Group providers

Grouping the results in a different way, and looking at student satisfaction in computer science providers in Russell Group (see Figure 4) and Million+ (see Figure 5) HEIs, a different picture emerges. Million+ providers have again Q15 as a candidate of high influence of student satisfaction, but Russell Group providers have a more consistent behaviour in Q02 (regarding the perceived interestingness of their subject).

AI-generated models

If we consider AI as a direct but more efficient replacement for analysis, we should expect similar results from an AI analysis of the same data.

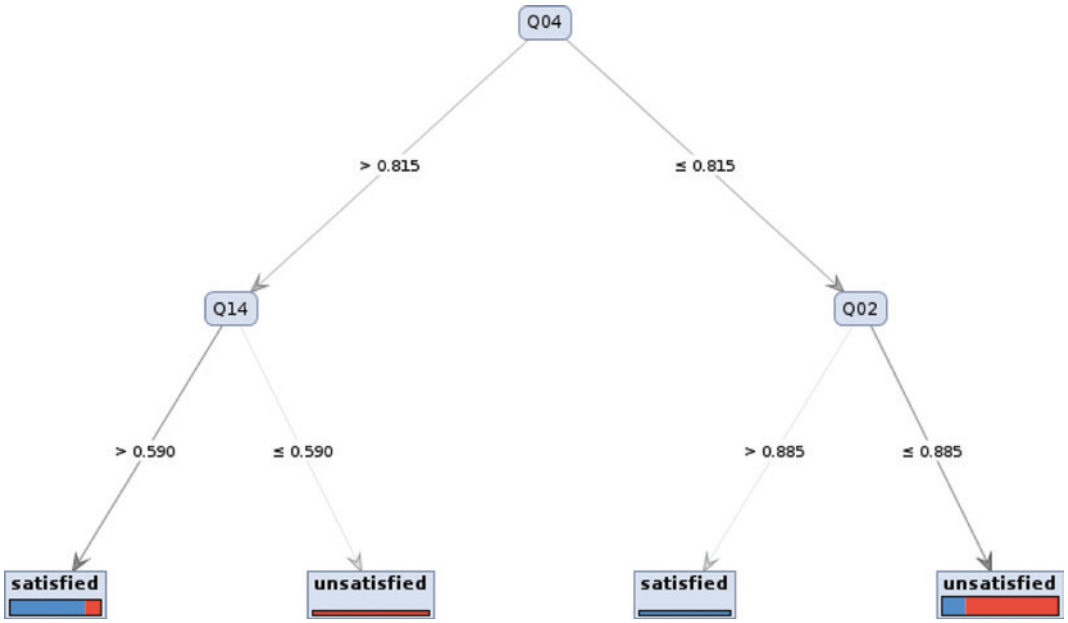

Figure 6: AI-generated tree model showing how Q22 satisfaction can be predicted from other NSS questions, in computer science, across all providers

The AI algorithm output (see Figure 6) disagrees with the conclusion of the manual analysis, but not significantly. It suggests that Q04 is the best question to split satisfied from unsatisfied students, followed by Q15 and Q01. Figure 6 suggests that students who are satisfied with Q04 but not Q01 will be mostly unsatisfied. However, students less satisfied with Q04 can still be satisfied overall if they are satisfied with Q15. This is not different to Figure 1, showing Q04 to be close to Q15 in importance.

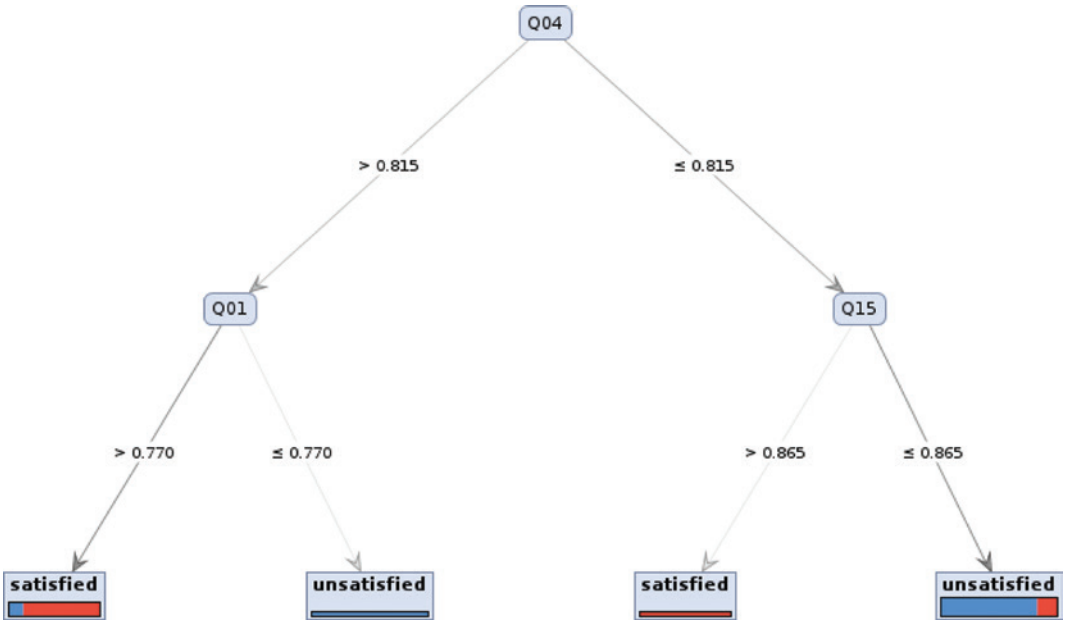

Figure 7: AI-generated tree model showing how Q22 satisfaction can be predicted from other NSS questions, in computer science, across the top half of providers by size

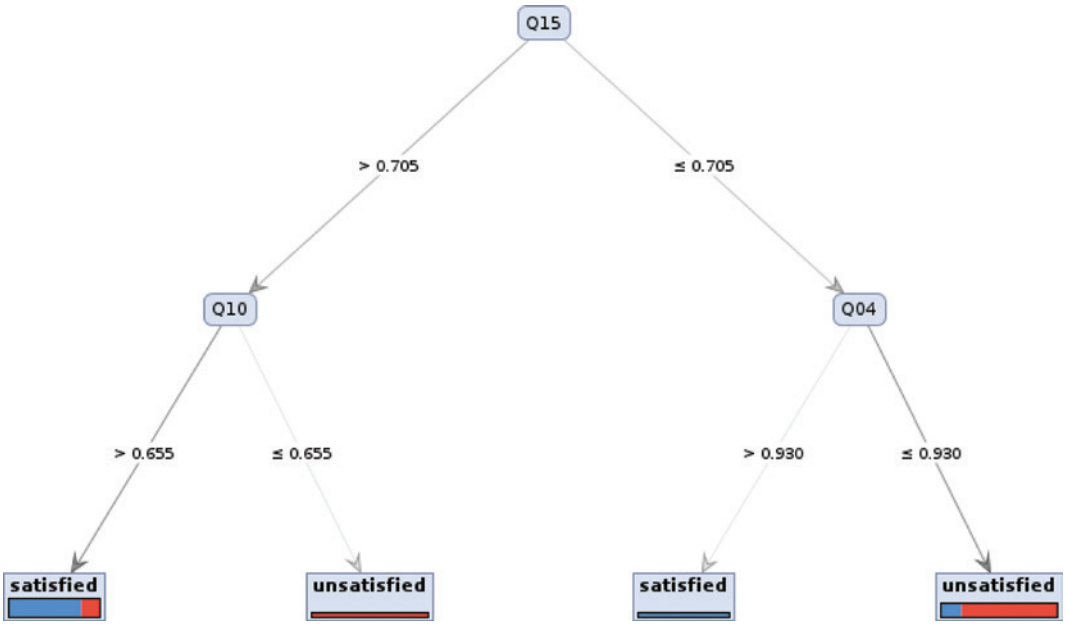

Figure 8: AI-generated tree model showing how Q22 satisfaction can be predicted from other NSS questions, in computer science, across the bottom half of providers by size

AI also states how confident it is that its model is accurate, and it has a 77.5 per cent confidence for Figure 6. Remaining close to the manual analysis, the AI algorithm insists that Q04 is the best choice when it looks at the larger providers of computer science (see Figure 7), with a confidence of 80.3 per cent, but it changes this to Q15 (see Figure 8) when looking at the smaller providers, with a slightly lower confidence of 76.0 per cent.

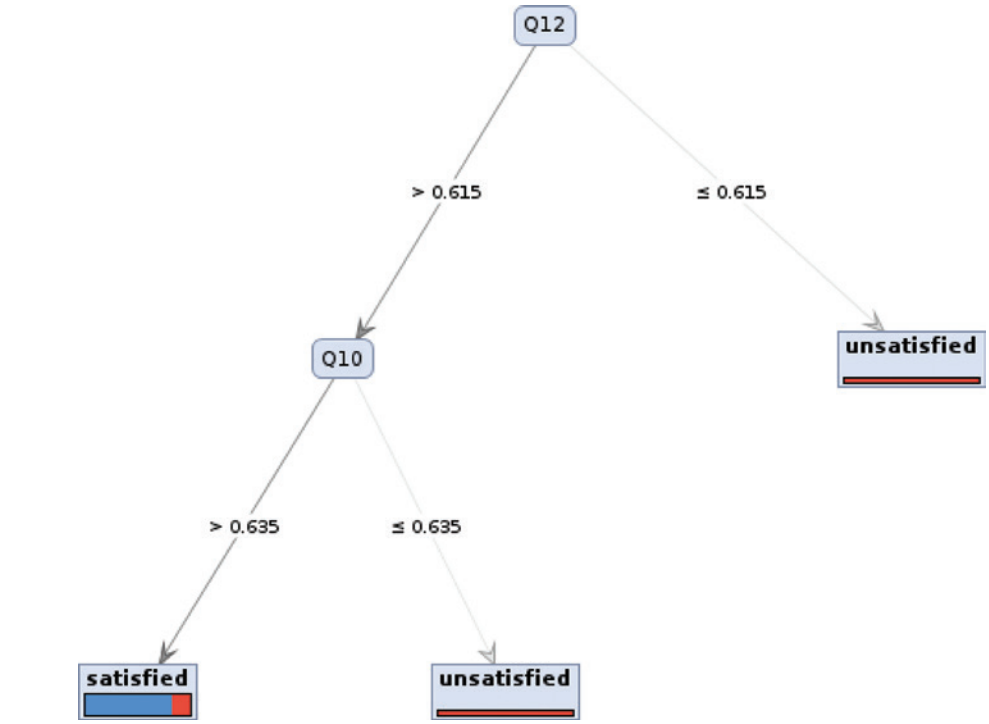

Figure 9: AI-generated tree model showing how Q22 satisfaction can be predicted from other NSS questions, in computer science, for Million+ Group providers

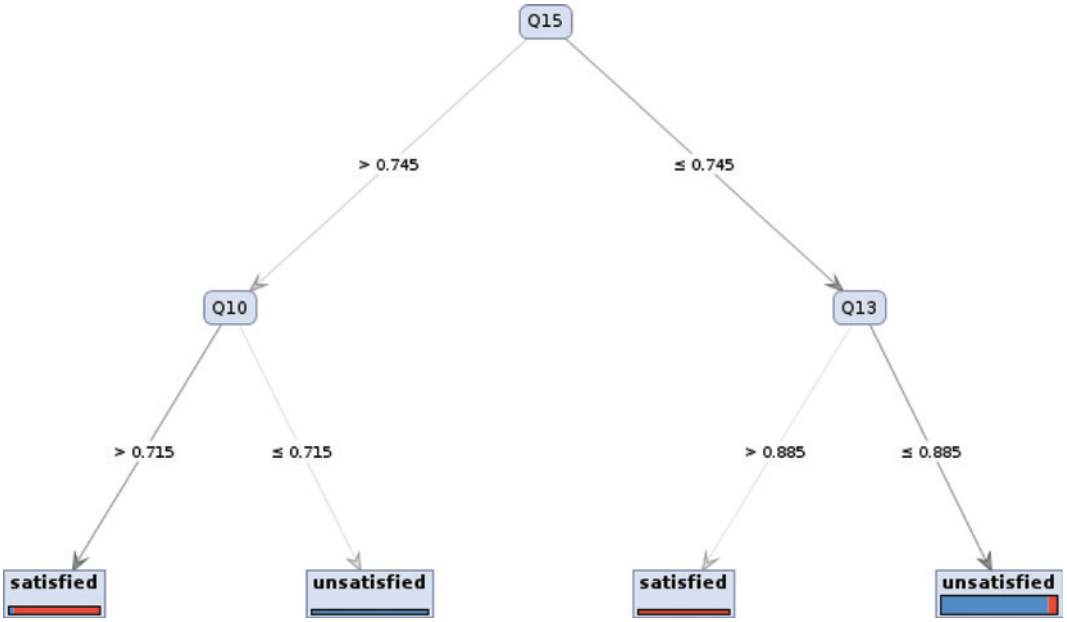

Figure 10: AI-generated tree model showing how Q22 satisfaction can be predicted from other NSS questions, in computer science, for Russell Group providers

Similarly, looking at the Million+ Group of providers (see Figure 9), the AI agrees that Q15 is the strongest candidate, raising its confidence to 81.6 per cent. However, with an even stronger confidence of 87.9 per cent, it points to Q12 (about good advice being available for study choices) for the Russell Group providers (see Figure 10), which the manual analysis would not rate so highly as a candidate.

In summary, the AI result that least agrees with the manual analysis is the one that the AI has the highest confidence in. This raises a question for HEIs: is a slow

and expensive manual analysis, which suggests prioritizing course organization and stimulating teaching, better than the faster and more cost-effective AI, which suggests prioritizing the quality of advice for study choices?

Running the same five sets of data through a Bayesian learner – a different AI algorithm using probabilities – AI gives predictive models with higher confidence (see Table 2).

Table 2: Confidence on their findings by different AI algorithms

Data set	Tree learner confidence (%)	Bayesian learner confidence (%)
All computer science	77.5	81.2
Top computer science by size	80.3	81.4
Bottom computer science by size	76.0	81.3
Russell Group computer science	81.6	81.4
Million+ Group computer science	87.9	82.1

However, like many AI algorithms, Bayesian learning does not produce models that can be inspected by users who wish to understand their reasoning.

Applying AI where there is no model

As a separate experiment, AI was used to investigate external factors that may influence student satisfaction across all subjects. As the NSS is not a complete survey of everything affecting student life, it would be natural for HEIs to expand the search into more university data to discover patterns of what influences student satisfaction. However, for this experiment, the data were chosen deliberately to have no known relationship to student satisfaction:

- the arbitrary numeric code assigned to each subject
- the length of the institution name
- the length of the subject name.

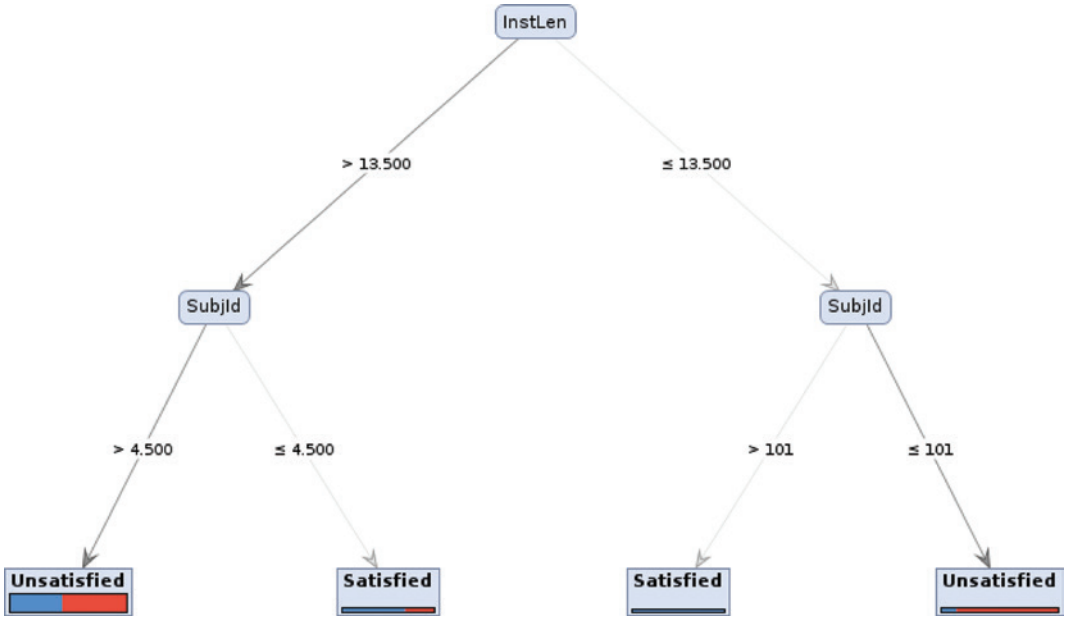

Figure 11: AI-generated tree model showing how Q22 satisfaction can be predicted from non-NSS data

The results shown in Figure 11 are not surprising: the algorithm's recommendation is that the road to student satisfaction is to shorten an institution's name to less than 13 characters. This misguided suggestion does not come with a very high confidence (54.7 per cent), while the Bayesian learner's modelling of the data is slightly more confident (56.1 per cent) about seeing a pattern where there is not one.

The presence of patterns in the National Student Survey

Concluding the experiments, Figure 12 shows a different form of pattern that can be found in the NSS. Using NSS scores rather than correlations, two HEIs are shown with NSS scores in all subjects (*y*-axis) and years (*x*-axis), on a blue to red gradient, with high satisfaction scores in red and low and no scores in blue. The values of note are those with the white background, which are the highest NSS satisfaction score that subject has achieved over the ten years. Scores are adjusted to show satisfaction from the full population rather than only those responding, explicitly excluding non-responders from those satisfied. The difference in how these peak values align is noteworthy. In the left HEI (HEI1), the different subject areas peak in different years without any signs of coordinated behaviour, while in the right HEI (HEI2), the different subject areas peak mainly in the last two years, suggesting an underlying institution-wide effort to improve the NSS scores.

Discussion

Educators made observations and decisions about learners' needs long before computers could access data. Educators also evolve and adapt from the experience of thousands of students. In turn, AI has shown its potential to analyse data at a massive scale with thousands of factors to support decisions, in a matter of minutes. Enhancing decision making with the insights of such a tool can be quite attractive – more so when decision makers are under pressure to improve results and efficiency.

Saving costs

The analysis of the heat-map models was informed by experience in the domain, reinforced through long analysis of subsets across institutions and subjects. This process took time (months over multiple years) and carries cost. The AI approach benefited from the data preparation already having been done, but it still offered undeniable advantages in speed (on this amount of data, AI can produce a new model, on a typical laptop, in seconds). AI also offers scalability: the volume of educational data can reach a point where manual analysis is prohibitive. This is where AI has been successful in taking over analysis in industry.

The ability to make informed decisions

The heat-maps demonstrate that patterns exist within NSS data and can persist across years. The finding that Q15 (about course organization) is the most important predictor of satisfaction of computer science students can help policy: the overall NSS satisfaction score in computer science can be improved by addressing student perception of course organization. This finding is also consistent with the literature, yet Bell and Brooks (2018) found that, despite this, some HEIs chose to address NSS satisfaction by prioritizing assessment and feedback; these score the lowest, but have much lower influence on overall satisfaction. This raises a question of how much capacity (and time) HEI decision makers have for deeper understanding of the data

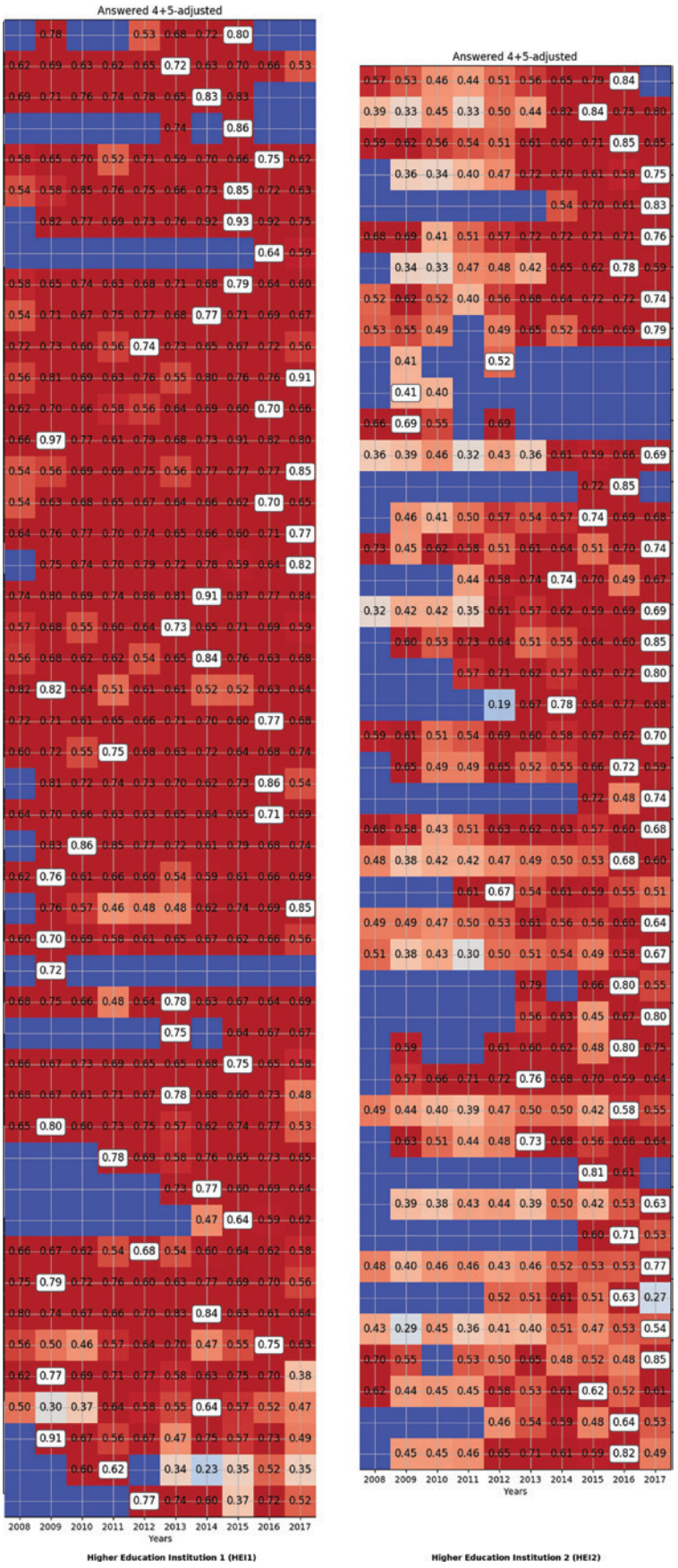

Figure 12: Comparison of when highest scores occur, per subject, in different HEIs

before making a strategic decision about improving overall satisfaction. AI cannot be held accountable for a wrong solution of an educational problem if its users do not have the capacity to fully understand the problem.

The lack of domain knowledge

Silver (2012) raises a problem with finding answers from algorithmic analysis of data: it can work very well when the data is clear and relationships are well understood – such as being able to distinguish an apple from an orange – but this is not the case where there is limited and noisy data, and when the understanding of the fundamental relationships is poor. Whereas a hundred people can provide a consistent description of apples and oranges, a hundred academics would not agree on a consistent description of how to raise student satisfaction. This is evident from the breadth of HEI strategies to improve NSS scores, but also from areas with sudden drops in NSS scores leading to new introspection and analysis of what the causes could be. AI works well with problems that are understood well. The full causes of student dissatisfaction are not just complex; they are also not well understood by those tasked to make decisions. AI cannot be held accountable for its analysis of an educational measure when there is no agreement about what is being measured.

Overfitting

Hawkins (2004) explains the problems of overfitting, when the model of the data is so closely tied to the specific data that it is useless for any future data. AI algorithms will need to be guided away from making decisions that keep them blind to how varied the data and their sources will be in real life (Parapadakis, 1999). Schaffer (1993) explains why any such choice is, in itself, introducing bias in the process. This is not an issue where the data are understood well, such as where an expert in apples will have a good appreciation of the variety of apple colours. It is an issue in areas where AI is used to find an unknown. Without having a universally agreed definition of a satisfied student, every HEI decision maker can bring a different bias on how to guide the algorithm, biasing it potentially towards a wrong finding. Furthermore, Ipsos MORI (2019) acknowledges the limits of understanding in HEIs by providing public advice on correct use of the data, as well as advice to students on how to avoid being inappropriately influenced. AI cannot be held accountable for fairness when bias can be introduced through the guidance from its users.

Questioning the reasoning

The better performance of the Bayesian learner is not surprising as it often outperforms tree learners in such tasks. However, like many of the better-performing AI algorithms, it will not explain its reasoning. Adadi and Berrada (2018: 52155) list the problems with black-box AI systems that produce findings without explanation, suggesting the need for explainable AI (XAI) and rightly pointing out that it 'is not enough to just explain the model, the user has to understand it'. This is also emphasized by Putnam and Conati (2019), specifically in education. For decisions to be made based on an AI analysis of the data, the decision maker needs to have not only transparency of why and how AI has come to its findings, but also the required understanding of what is being modelled, so as to ask the right questions to gain full understanding of these findings. XAI is not a minor academic concept, but lies behind a real need in AI recognized by the Defense Advanced Research Projects Agency (DARPA) (see Gunning, 2017), as well as in the European Union's General Data Protection

Regulation's (GDPR) 'right to explanation' (Goodman and Flaxman, 2017). AI cannot be held accountable for its flaws if its reasoning can never be audited.

Apophenia

The last AI experiment showed how AI can advise a decision maker to change the name of the HEI to improve student satisfaction. The experiment was deliberately extreme, but it is illustrative of the problem of apophenia (Goldfarb and King, 2016). In the same way that we may look at clouds and see faces, AI will find models that are not there: even random data will occasionally contain patterns. Although the low confidence can be used as a threshold to filter out nonsense suggestions, it is important to reiterate that education is not like typical applications of AI; the underlying models are not understood well enough for HEIs to agree on what constitutes good student satisfaction. A decision maker faced with advice to shorten the name of the HEI will rightly be suspicious, but what if AI suggests, for example, with 66.5 per cent confidence, that student satisfaction can be raised by offering summer internships abroad, abolishing morning lectures, or making the use of a specific technology compulsory for all lessons? All would sound plausible and it would not be possible for the decision maker to know if AI has found something really useful or found a pattern where there really was none. AI cannot be held accountable for accuracy if its users think even random findings sound plausible.

Reproducibility of the experiment

The success of AI is not only in identifying patterns that are clear in the data, but also based on the assumption that such patterns are persistent. When AI is being trained by looking at successive examples of apples and oranges, it is trained on the assumption that oranges will not alter their appearance over the time they are being looked at. Yet this assumption cannot exist when analysing student satisfaction data. As would be expected, HEIs are studying the data each year, and they adapt their policies so as to improve student satisfaction. Figure 12 shows that institutions can make decisions that influence NSS results. Changes such as this do not appear to influence patterns discussed earlier, as the data comes from multiple HEIs, while an HEI strategy influences just one HEI. But that does not reduce the importance of this assumption in AI. UK-wide evaluations of HEI research have existed for far longer than the NSS; while reviewing the Research Excellence Framework (REF), Stern (2016: 12) noted about HEIs that 'we are wary of tactics designed to maximize REF performance that may not be harmonious with the longer-term fostering of quality research and staff development in the sector as a whole'. A decision maker who uses AI on past data to make a decision that changes future data cannot then use both sets with AI. AI algorithms have been successful on problems where new data are largely consistent with previous data. AI algorithms are not designed to recognize oranges that adapt each year so they can look more like apples, and AI cannot be held accountable for guidance that comes from users who can tweak the data to get the guidance they prefer.

Conclusions

The NSS contains useful and detectable patterns of student behaviour. This suggests that education in general can have detectable patterns of behaviour. Analysing and identifying such patterns – where they exist – can help decision making, and better decisions can lead to benefits for learners. The scale and complexity of data involved means an analysis done by traditional means can carry a significant annual cost. The

successes of AI in industry and its cost effectiveness make it an attractive alternative to deploy on educational data to help in a variety of problem areas.

Issues affecting the proper deployment of AI algorithms have been recognized before. Cohen and Howe (1988) argue for the need for proper evaluation stages of AI, and three decades later, Lipton and Steinhardt (2018) similarly argue that the recent rapid success of machine learning has 'troubling trends' that lead to flaws in scholarship. AI is mature enough to be trusted to support decisions by analysing large volumes of data in certain domains, but its successes have been with data that are well understood (such as detecting malware), data where it is easy to evaluate the algorithm's success (such as face recognition) and data where anything better than a random guess brings benefits (such as predicting customer behaviour). Education fits none of these categories. It is not a question of whether education is ready for AI, but a question of when (and possibly if) AI will be mature enough to be helpful to education.

Strategic decisions where universities address NSS overall dissatisfaction by prioritizing improvements in assessment and feedback, and where decisions are made by comparing subject scores to the institution averages, demonstrate limitations of understanding the data at high levels of HEIs, before making decisions about student needs. A tool that 'provides answers' could be adopted as a welcome time saver without realizing either the extended understanding of the domain that AI assumes or how relatively limited is our understanding of factors affecting the data, compared with domains where AI has been successful.

A research experiment to evaluate student satisfaction that is based on no understanding of the data would never be evaluated by an ethics committee on the basis of scalability, speed and cost. Similarly, to judge if AI can effectively replace (in whole or part) the educational researcher, scalability, speed and cost are misleading criteria. Decision makers may never have the capacity or afford the time to understand fully all the complexities of a large-scale educational problem before using AI. This is not an issue of prioritizing what makes pedagogical sense, but an issue of appreciating that, in many educational problems, our pedagogical sense is not yet good enough for AI. The issue here is not one of enshrining in law the right of algorithmic accountability, but one of acknowledging that algorithmic accountability may not ever be possible in a domain that is not sufficiently understood, on metrics where there is disagreement about what they measure, by users who can neither examine the decision making in the algorithm nor have the capacity to fully understand the problem being analysed. In the areas of education with the most debate between researchers, AI may provide some helpful insights, but it cannot support decision making with algorithmic accountability.

Notes on the contributor

Dimitris Parapadakis is a principal lecturer at the University of Westminster. Since 1991, he has developed and led projects in AI. As an academic he has held, since 2000, various roles at different HEI levels overseeing quality assurance and quality enhancement. He is the head of the university's Cyber-Security Research Group and a co-convenor of the Artificial and Human Intelligence Special Interest Group in the British Educational Research Association.

References

Adadi, A. and Berrada, M. (2018) 'Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)'. *IEEE Access*, 6, 52138–60.

Agnew, S., Cameron-Agnew, T., Lau, A. and Walker, S. (2016) 'What business school characteristics are correlated with more favourable National Student Survey (NSS) rankings?'. *International Journal of Management Education*, 14 (3), 219–27. http://dx.doi.org/10.1016/j.ijme.2016.05.001.

Bell, A.R. and Brooks, C. (2018) 'What makes students satisfied? A discussion and analysis of the UK's National Student Survey'. *Journal of Further and Higher Education*, 42 (8), 1118–42. https://doi.org/10.1080/0309877X.2017.1349886.

Boehler, M.L., Rogers, D.A., Schwind, C.J., Mayforth, R., Quin, J., Williams, R.G. and Dunnington, G. (2006) 'An investigation of medical student reactions to feedback: A randomised controlled trial'. *Medical Education*, 40 (8), 746–9. https://doi.org/10.1111/j.1365-2929.2006.02503.x.

Buckley, A. (2012) *Making It Count: Reflecting on the National Student Survey in the process of enhancement*. York: Higher Education Academy.

Cheng, J.H.S. and Marsh, H.W. (2010) 'National Student Survey: Are differences between universities and courses reliable and meaningful?'. *Oxford Review of Education*, 36 (6), 693–712. https://doi.org/10.1080/03054985.2010.491179.

Cohen, P.R. and Howe, A.E. (1988) 'How evaluation guides AI research'. *AI Magazine*, 9 (4), 35–43.

DfE (Department for Education) (2017) *Teaching Excellence and Student Outcomes Framework Specification*. London: Department for Education.

Fielding, A., Dunleavy, P.J. and Langan, A.M. (2010) 'Interpreting context to the UK's National Student (Satisfaction) Survey data for science subjects'. *Journal of Further and Higher Education*, 34 (3), 347–68. https://doi.org/10.1080/0309877X.2010.484054.

Frankham, J. (2015) 'Much ado about something: The effects of the National Student Survey on higher education'. BERA blog, 11 August. Online. https://tinyurl.com/yx6lr77p (accessed 3 April 2020).

Goldfarb, B. and King, A.A. (2016) 'Scientific apophenia in strategic management research: Significance tests and mistaken inference'. *Strategic Management Journal*, 37 (1), 167–76. https://doi.org/10.1002/smj.2459.

Goodman, B. and Flaxman, S. (2017) 'European Union regulations on algorithmic decision making and a "right to explanation"'. *AI Magazine*, 38 (3), 50–7. https://doi.org/10.1609/aimag.v38i3.2741.

Gunn, A. (2018) 'Metrics and methodologies for measuring teaching quality in higher education: Developing the Teaching Excellence Framework (TEF)'. *Educational Review*, 70 (2), 129–48. https://doi.org/10.1080/00131911.2017.1410106.

Gunning, D. (2017) 'Explainable artificial intelligence (XAI)'. Defense Advanced Research Projects Agency (DARPA). Online. https://tinyurl.com/y7p4ex65 (accessed 20 April 2020).

Haenlein, M. and Kaplan, A. (2019) 'A brief history of artificial intelligence: On the past, present, and future of artificial intelligence'. *California Management Review*, 61 (4), 5–14. https://doi.org/10.1177%2F0008125619864925.

Hall, W. and Pesenti, J. (2017) *Growing the Artificial Intelligence Industry in the UK*. London: Department for Digital, Culture, Media and Sport and Department for Business, Energy and Industrial Strategy.

Hawkins, D.M. (2004) 'The problem of overfitting'. *Journal of Chemical Information and Computer Sciences*, 44 (1), 1–12. https://doi.org/10.1021/ci0342472.

Hinojo-Lucena, F.-J., Aznar-Díaz, I., Cáceres-Reche, M.P. and Romero-Rodríguez, J.-M. (2019) 'Artificial intelligence in higher education: A bibliometric study on its impact in the scientific literature'. *Education Sciences*, 9 (1), 1–9. https://doi.org/10.3390/educsci9010051.

House of Commons Education Committee (2018) *Value for Money in Higher Education: Seventh report of session 2017–19*. London: House of Commons.

Ipsos MORI (2019) *National Student Survey 2020 Good Practice Guide: Marketing and promoting the National Student Survey*. Online. https://tinyurl.com/ya7omyfe (accessed 20 April 2020).

Khan, A., Baharudin, B., Lee, L.H. and Khan, K. (2010) 'A review of machine learning algorithms for text-documents classification'. *Journal of Advances in Information Technology*, 1 (1), 4–20. https://doi.org/10.4304/jait.1.1.4-20.

Langan, A.M., Dunleavy, P. and Fielding, A. (2013) 'Applying models to national surveys of undergraduate science students: What affects ratings of satisfaction?'. *Education Sciences*, 3 (2), 193–207. https://doi.org/10.3390/educsci3020193.

Lipton, Z.C. and Steinhardt, J. (2018) 'Troubling trends in machine learning scholarship'. *arXiv*, 1–15. Online. https://arxiv.org/pdf/1807.03341.pdf (accessed 3 April 2020).

McCord, M. and Chuah, M. (2011) 'Spam detection on Twitter using traditional classifiers'. In Alcaraz Calero, J.M., Yang, L.T., Gómez Mármol, F., García Villalba, L.J., Li, A.X. and Wang, Y. (eds) *Autonomic and Trusted Computing: 8th International Conference, ATC 2011, Banff, Canada, September 2–4, 2011: Proceedings*. Berlin: Springer, 175–86.

Naftulin, D.H., Ware, J.E. and Donnelly, F.A. (1973) 'The Doctor Fox Lecture: A paradigm of educational seduction'. *Journal of Medical Education*, 48, 630–5.

OfS (Office for Students) (2018) *Office for Students Strategy 2018 to 2021*. London: Office for Students.

Parapadakis, D. (1999) 'The Hydra system: A machine learning system for multiple, impure, sources'. In Mohammadian, M. (ed.) *Computational Intelligence for Modelling, Control and Automation: Intelligent image processing, data analysis and information retrieval*. Amsterdam: IOS Press, 183–8.

Parapadakis, D. and Baldwin, M. (2018) 'TEF and NSS: What makes computer science students so unhappy?'. Paper presented at the BERA Conference, Newcastle upon Tyne, 11 September.

Putnam, V. and Conati, C. (2019) 'Exploring the need for explainable artificial intelligence (XAI) in intelligent tutoring systems (ITS)'. Paper presented at the IUI Workshops'19, Los Angeles, 20 March 2019.

Sabri, D. (2013) 'Student evaluations of teaching as "fact-totems": The case of the UK National Student Survey'. *Sociological Research Online*, 18 (4), 1–10. https://doi.org/10.5153%2Fsro.3136.

Schaffer, C. (1993) 'Overfitting avoidance as bias'. *Machine Learning*, 10 (2), 153–78.

Silver, N. (2012) *The Signal and the Noise: The art and science of prediction*. London: Allen Lane.

Stern, N. (2016) *Building on Success and Learning from Experience: An independent review of the Research Excellence Framework*. London: Department for Business, Energy and Industrial Strategy.

Sun, Y., Chen, Y., Wang, X. and Tang, X. (2014) 'Deep learning face representation by joint identification-verification'. Online. https://tinyurl.com/ycq6ghtj (accessed 20 April 2020).

Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G. and Chatzisavvas, K.C. (2015) 'A comparison of machine learning techniques for customer churn prediction'. *Simulation Modelling Practice and Theory*, 55, 1–9. http://dx.doi.org/10.1016/j.simpat.2015.03.003.

Yorke, M., Orr, S. and Blair, B. (2014) 'Hit by a perfect storm? Art and design in the National Student Survey'. *Studies in Higher Education*, 39 (10), 1788–810. https://doi.org/10.1080/03075079.2013.806465.

Zawacki-Richter, O., Marín, V.I., Bond, M. and Gouverneur, F. (2019) 'Systematic review of research on artificial intelligence applications in higher education – where are the educators?'. *International Journal of Educational Technology in Higher Education*, 16, Article 39, 1–27. https://doi.org/10.1186/s41239-019-0171-0.

Zawbaa, H.M., Hazman, M., Abbass, M. and Hassanien, A.E. (2014) 'Automatic fruit classification using random forest algorithm'. In *2014 14th International Conference on Hybrid Intelligent Systems*. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 164–8.